UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Working Paper A10/02

Applications and Policy

Fundamental Principles In Drawing Inference From Sequence Analysis

Tom King

Abstract

ABSTRACT

Individual life courses are dynamic and can be represented as a sequence of states for some portion of their experiences. More generally, study of such sequences has been made in many fields around social science; for example, sociology, linguistics, psychology, and the conceptualisation of subjects progressing through a sequence of states is common. However, many models and sets of data allow only for the treatment of aggregates or transitions, rather than interpreting whole sequences. The temporal aspect of the analysis is fundamental to any inference about the evolution of the subjects but assumptions about time are not normally made explicit. Moreover, without a clear idea of what sequences look like, it is impossible to determine when something is not seen whether it was not actually there. Some principles are proposed which link the ideas of sequences, hypothesis, analytical framework, categorisation and representation; each one being underpinned by the consideration of time. To make inferences about sequences, one needs to: understand what these sequences represent; the hypothesis and assumptions that can be derived about sequences; identify the categories within the sequences; and data representation at each stage. These ideas are obvious in themselves but they are interlinked, imposing restrictions on each other and on the inferences which can be drawn.

Fundamental Principles in Drawing Inference from Sequence Analysis

TOM KING

Division of Social Statistics and Southampton Statistical Sciences Research Institute

University of Southampton

tom.king@soton.ac.uk

ABSTRACT

Individual life courses are dynamic and can be represented as a sequence of states for some portion of their experiences. More generally, study of such sequences has been made in many fields around social science; for example, sociology, linguistics, psychology, and the conceptualisation of subjects progressing through a sequence of states is common. However, many models and sets of data allow only for the treatment of aggregates or transitions, rather than interpreting whole sequences. The temporal aspect of the analysis is fundamental to any inference about the evolution of the subjects but assumptions about time are not normally made explicit. Moreover, without a clear idea of what sequences look like, it is impossible to determine when something is not seen whether it was not actually there. Some principles are proposed which link the ideas of sequences, hypothesis, analytical framework, categorisation and representation; each one being underpinned by the consideration of time. To make inferences about sequences, one needs to: understand what these sequences represent; the hypothesis and assumptions that can be derived about sequences; identify the categories within the sequences; and data representation at each stage. These ideas are obvious in themselves but they are interlinked, imposing restrictions on each other and on the inferences which can be drawn.

KEY WORDS: *sequence analysis, dissimilarity algorithms, stochastic models, event histories, inference, visualisation, clustering*

1. Structure

The relation between data and theory is the basis of science; the process of validating theory by examining patterns in data, we call inference (Hume 2000). Data occur in many forms which require different considerations to be understood: sequence analysis is the study of the evolution over time of some characteristic of some subjects. The focus of analysis to make inferences from these recorded evolutions about the process which generated them by quantifying patterns in and amongst the sequences themselves. For example, these are often people and the evolutions are their life courses, process generating these being their own consciousness and the nature of their external environment (Levy et al. 2005).

The complexity of processes that may be governing the evolution of one individual is enormous so that there is a valid argument for treating individuals separately. However, a simplification through which inferences are made about these processes would be deeply valuable: there are many applications which track a qualitative evolution over time and many disciplines where such methods can be applied. Hence we proceed here to develop a framework through which inferences might be made, given due allowance for the nature of the theory being studied, without making undue assumptions about the aims or concerns of particular research.

Our intention is not to produce a method which is prescriptive, which is not particularly feasible at this stage (see §2), and may well be undesirably limiting given the interdisciplinary nature of the current applications (Feyerabend 1988). Instead we evince some principles to guide an analysis of sequence data and, in order to achieve this, describe the key elements of such an analysis. The principles themselves are manifest as the dependencies between the elements (see §8) and indeed they exhibit the recursive structure intrinsic to many research methods, such as case studies. Although we do review the research so far published on such matters, and the success these studies have had in making inferences about the theory they sought to study, the focus is on the epistemic rather than the substantive. So we derive some ideas which should be considered iteratively to guide thinking rather than obstruct it, whilst leaving the essential component of time central to all of the interrelations.

Although there is some logical structure to the text which considers each of the elements in turn, the nature of the dependencies between these sections makes it highly non-linear in actuality. Moreover as the whole interrelation is complex and in some cases subtle, the cross-references given between the sections are very general but also try to describe the reference being made instead referring the reader to a specific paragraph as the extent of this would be frustrating on first reading. Although this will result in some repetition, this will be more conducive to understanding, particularly due to the iterative nature of the implementation, and the interdisciplinary nature of the context which gives several different perspectives to the same problems.

We reveal a situation similar to that observed by John Goldthorpe, albeit on another topic, that of the clash between qualitative and quantitative study in social science:

"while the issues caught up in the protracted and complex exchanges that have occurred include ones of major importance, the form that the debate has taken has not been especially helpful in highlighting just what those issues are, nor yet in pointing to ways in which they may be effectively addressed."

(Goldthorpe 2007, 7)

The clash in our case is between the analytical frameworks described in §5, and while we do not intend to denigrate the approach of analysis for its own sake, indeed Chomsky (1957) eloquently describes the utility of analysis, there are clearly some situations which are more appropriate to one rather than the other. So, we hope in the concluding section to draw together the important barriers to inference and so the necessary steps for further progress in sequence analyses.

The main thrust of this paper is pragmatic, despite its rather philosophical interludes. Hence the sections not yet referred to deal with practical issues of the nature of sequences (§3), the hypotheses they seek to evaluate (§4), the categorisation of the states they sequence (§6) and the modes used to represent the sequences and the results of their analysis (§7). They follow a general approach of defining the element, relating it to the other elements and exemplifying the importance of this in published work. These sections derive the various principles forming §8, and also indicate various strengths and weaknesses which form the conclusion of this paper.

2. Genesis

Science has always studied empirical data, whether for understanding or prediction, and these data relate to observations of an equivalent closed system. In comparing what was measured in separate observations of the system, one seeks a relationship which explains our observations, perhaps through the different circumstances involved. We try to unpick resemblances from causal and contiguous relationships (Hume 2000:16), whether these are internal or external similarities we study.

One of the principal factors differentiating one observation of a system from another is the time at which is made. Indeed time is seen to be important not simply in the study of autonomous systems by the time spent in the observation, but also for its psychological importance to humans (Poincaré 1913): its subjective interpretation. Thus science has great interest in studying the evolution of the same system over time, as do the scientists involved in the study.

Measurement is of considerable importance to scientific endeavour (Feyerabend 1988, 2nd) and indeed science and measurement have advanced hand in hand (Kuhn 1996). Thus large quantities of observations and their associated measurement systems have accrued complex datasets of the relations between them. Thence the scientist has sought to simplify the system under study to one where the dataset is simple enough to be comprehended in its entirety, i.e. one seeks to refine the design of one's experiment to study specific patterns and subsets. Where this is not possible, as it so often is not in the case of social science, one seeks to model the dataset by drawing out the variation relevant to the theory being advanced whilst making assumptions about the residual relationships in the data. So models are often used to circumvent difficulties presented by measurement and the design by which observations are made, for example this may be the only way of studying macroeconomics.

Models for longitudinal data are versatile to the practical problems of testing hypotheses, even for fairly complex systems, and where these data, possibly discrete, are linearly ordered, there are satisfactory approaches for assessment of the validity of assumptions made by the models. However, where longitudinal data are categorical and representing them on an ordinal scale is not satisfactory, not an unusual situation

in the social sciences, there is a considerable conflict between modellers and theorists. This conflict is complex, and often personal, so it is best understood from an historical perspective of its development. Indeed, for analysis of longitudinal categorical data in social science, there is no consensus despite much work, as is discussed in later sections. The evolution of the methods will be traced in the remainder of this section.

In the fifty years that the analysis of sequences has been an active interest, the methods available for their analysis have developed and diversified. Indeed, one of Chomsky's (1957) qualifications of his pioneering work was that he should find the shortcomings of the models available with regard to the prediction of semantic structure. He did not assert that his analysis would give a complete understanding of the data but that it would elucidate the structure of the data which did not fit the model. Thus many researchers have led the way with methodological developments necessitated by the limited tools available. Indeed the current consensus is that there are weaknesses to the methods in use, giving strong motivation for a comprehensive review of their efficacy.

Many examples of sequences come from the observation of systems over time, being composed of events, transitions or whatever is appropriate (see §3 for discussion of sequences and a description of different conceptualisations). Although there are studies of other kinds of sequences, particularly arising in biological science, and most notably DNA, the first analytical considerations of sequences focussed on the conscious expressions of living organisms. Indeed, the analysis of sequences has been driven by the conceptual interest of such data, and many applications have involved empirical work, each in its separate field. Specifically, initial studies focussed on patterns in linguistics (Chomsky 1957), and animal behaviour (Altman 1965). Since that time, the methods have been adopted more generally in sociological applications (e.g. Billari and Piccarreta 2005;Blair-Loy 1999) but also in disparate fields such as geography (Wilson 1998), psychology (Bakeman and Quera 1995), and computer science (Iske 2009).

The continuing motivation for analysis of sequences has been the availability of data, with the obvious idea that analysis can aid the understanding of its generation. Inferences should be possible from this data which elucidate different things to other

analyses: the experiences of the individuals concerned. However, just as in all statistical modelling, there has been a clash between the assumptions of the methods used and the theorised structure of the data. Thus there are not so much competing methods as contrasting analytical frameworks which have developed originally in different fields and from diverse epistemic standpoints (see §5). This empirical nature of the work has meant that sequence analysis has developed separately in several disciplines and is only recently seeing an interdisciplinary perspective (Levy, Ghisletta, Le Goff, Spini, and Widmer 2005). This motivation from the data has also kept the methods away from statistical literature and meant that take up in disciplines far removed from social science has been delayed and unresponsive to new developments. There has also been a lot of attention given to some issues which are of little importance, for example algorithms do not to treat sequences symmetrically (have the distance between two the same in both directions) or even be metrics (multi-dimensional scaling of non-metric dissimilarities is perfectly possible (Cox and Cox 2001)), indeed some social processes would be at odds with these assumptions.

There was a promising beginning with an interdisciplinary compilation (Sankoff and Kruskal 1983) and the motivation of using dynamic programming to make study of life histories less subjective (Kruskal 1983). The slow development that has taken place has been associated with individuals in some disciplines championing their method which is natural when others do not take this up. However, this has the unfortunate consequence of the method being associated with the individual rather than the problem. Specifically in social science, discussion has focussed on Abbott's methods (Hollister 2009;Levine 2000) rather than how they might be improved; also documented in part of his own book (Abbott 2001) is the problem of an individual struggle with the difficulties he had, such as in approaching leaders in the field (telephoning Joseph Kruskal is likened to "calling God"). Some new work (Elzinga and Liefbroer 2007;Halpin 2010;Piccarreta and Lior 2010;Pollock 2007;Wilson 2006) suggests this is changing as people customise to their own problem and try to address specific issues.

Recently the methodology has become more accessible, particularly from a computational standpoint. All of the methods used come up against the problem of the amount of data and complexity of the analysis which led to the need to program

specially. DNA benefited from its invariant nature and the simple variation involved and so has a lot of software developed solely for its examination. Several models and algorithms have been available for some time but their implementation in mainstream statistical packages is a big step, which few have yet made. The recent explosion of publications using optimal matching to analyse sequences almost certainly owes more to its availability in STATA (Brzinsky-Fay, Kohler, and Luniak 2006) than any particular endorsement as a viable methodology. Indeed, aspects of best practice evident in the published literature are not implemented and still present barriers to research. Indeed stochastic models continue to advance from a theoretical perspective but have weaker links with the fields of application rather than specific projects (see Green, Lid Hjort, and Richardson 2003).

The whole concept of disciplinarity has been a substantial barrier to the effective use of sequence analysis not least because disciplines have had their own epistemic traditions. The semi-parametric nature of many methods has meant they do not find acknowledgement in the more positivist traditions of quantitative social science (c.f. Feyerabend 1988, 2nd;Goldthorpe 2007, 2nd), and so there has been little outlet for their publication. However, when they are published and read by outsiders, their epistemological foundation can be taken for granted; indeed it has taken interdisciplinary work to establish the depth of this problem and the need for a comprehensive consideration (Levy, Ghisletta, Le Goff, Spini, and Widmer 2005). What is needed is a clear and explicit framework linking together all of the ideas, issues, assumptions and problems associated with the analysis of sequences so that well informed decisions can be made an researchers can use the best tools available to answer their questions.

A fundamental problem in the analysis of sequences is that simply thinking about the processes as generating sequences is unnatural to those used to focussing on transitions or durations. This runs deeper than differences in existing traditions: it is close to what Feyerabend called the establishment of method. The holistic view championed by Abbott needs a holistic epistemology but it is clear enough that ‘we are not thinking sequentially’ at this point. Thus it is fundamental to consider in any research problem in this area exactly what is meant by a sequence, what it means

substantively in the theory, what may differentiate such sequences, how they are composed, and finally how they may be usefully represented.

3. Sequences

The evolution over time of a system is the intention of sequence analysis. To make a quantitative analysis of such evolution, even if not a fully parametric one, requires a definition of what data look like and how differences between data points can be characterised. Thus we must establish how such systems can vary, both over time and between individuals: it is the patterns in these variations which describe the evolution.

Sequences are made up of states and associated timings. Each individual is classified as having particular states at particular times and their sequence can be derived from this information. How best to describe what these states are is less straightforward than is desirable in complete generality. Instead we refer to different conceptions of state, e.g. stage, transition, event (Levy 2005) and the psychological timing of these states (Lesnard 2006). What is actually used in an individual analysis is dependent on the theory under study, as will the particular categorisation of the state variable. A sequence will normally use only one state variable and so each sequence can be represented as an ordered list of states, each with a corresponding timing. So a sequence is a function associated with an individual, mapping some portion of time to some state space.

We refer to individuals or subjects, and each one has a sequence associated with it. In most published work on sequence analysis, an individual refers to a single person, or possibly a pair of people or family unit. However, there are other examples where each individual is a different social construct such as a business, or a string of syntax, either musical or prose. As these also represent evolution of a system over time, they are included in the concept of a sequence; the consideration of DNA is quite different. Although DNA is commonly related as an example of sequences, it is a sequence unordered in time: evolution of the state of a subject over time represents something very different to progress along a strand. A sequence may be many things, but the experience of time is fundamental to this work and also to the process of inference.

The definition above is deeply abstract in an effort to preserve its generality. However, it is more helpful to explore some features that many sequences will have. They can be composed of episodes in one state followed by another, and each episode may be composed of several observations of the same state. We describe the lowest level of observation as the elements of the sequence. When the timing of the observations is sporadic, it may be more natural to consider episodes rather than elements. Also, for the purposes of the investigation of the theory, it may be more natural to consider one or the other, and to put more or less importance on the timing of the events, their temporal ordering, or simple existence. Birth histories may progress obviously through the first, second and third, the interest being the timings, whereas for social events like marriage, parenthood, house ownership, employment, the ordering may be of much more interest than timing.

Sequences represent the evolution of individual subjects over time and can therefore be a simple representation of particular features of each subject. However, any intention to make an inference or consider a theory will require comparison between the sequences followed by different subjects. Even descriptive analysis will benefit from some representation of whole sequences in relation to each other; more advanced work may want filtering, ordering or comparison in several dimensions (Piccarreta and Lior 2010, 173). The most natural form of these will be dependent on the nature of the sequences including how they are related to time (see §7 for more details).

The data social sequences come from often simply imply further problems: surveys have weights, clustering and attrition to deal with. There is not knowledge of how informative these problems may be, although sequence analysis may add to the understanding of some of them. These data are often complex, perhaps multidimensional in nature, and often include other time-constant, subject-level demographic variables (e.g. sex or country of birth). This does not make for promising sources indeed successful analyses have generally used cohort data. However, the data exist as do the theories so to dismiss their use out of hand is a little cavalier, if not irresponsible: certainly past analyses provide some ideas for improvements.

Examples of sequence representation include a continuous curve tracing a path over time, or a sequence of codes, like a sequence of notes from a musical score. Visual comparison is immediately complicated by the possibilities available to the curvature of the trajectory, the demarcation of the space, and the separation of the curves; or the alignment of sequences in the discrete case. Indeed it is these considerations which have driven comparisons in the holistic analytical framework but even these speculations immediately make it clear that long sequences or some adjustments will be necessary to cope with measurement error (see §5 for further discussion).

While a sequence tries to represent the whole experience of a single subject, it also attempts to make it more accessible. It reduces to some state space of what can be experienced to develop a proxy in categories for the subject itself (indeed the categorisation is a subtle process, see §6). While this must sacrifice some information, it allows for the drawing of comparisons between subjects, which then allows for treatment of theory with the same approach: we are looking for similarities between subjects and sequences (Kruskal 1983; see §4). It is also very difficult to ensure that it is equivalent for all the subjects under study throughout their sequences. Thus categorisation may be specific to one study, variable over time, or multidimensional in nature. The key is that it is able to embody the variation in evolution of subjects over their whole sequences in relation to the theory under study.

Research in social science often studies particular events or transitions between stages in the lives of people. These are often marked by certain rites of passage for the transition, which are often worthy of study in their own right, but these can distract from the actual transition made between stages. Specifically, the timing of recognition may be logical in that it comes when the transition is in some sense irreversible (e.g. marriage requires divorce but an engagement may be broken off) but this may not be how it is encapsulated in the theory. Indeed a substantial barrier to using sequences in analysis has been making links between theory and data; simply thinking sequentially is a new approach. This is complicated by data collection reflecting the hard transitions not the more abstract lifestyle changes involved in life course theory (Levy, Ghisletta, Le Goff, Spini, and Widmer 2005). Thus the timing system used for the changes of state experienced by subjects, forming their sequences should be

derived from the theory and hypothesis, not reliant solely on data conventions used elsewhere.

The novel nature of sequence analysis, and the complexity of social data and theory, implies that useful precedents for a given analysis are lacking. Thus relating the process and variation expected from the theory to appropriate sequence data for subjects is a conceptual challenge, especially when using a holistic analytical framework. It is safe therefore to rely on a coarse categorisation over long sequences and represent them without processing (c.f. Clark, Deurloo, and Dieleman 2003). Actually conceptualising the data as sequences rather than component elements, stages and transitions is essential for meaningful analysis but very rare in practice. Making inferences about a holistic theory will require sequences which embody the holistic nature of the data, and therefore a hypothesis about how they should do this.

4. Hypotheses

The problem of the scientific method has concerned many; indeed criticism of the perceived dominance of certain approaches has become a substantial endeavour. This has given existing methods an entrenched position despite the recognition of the possibility of a paradigm rather than simple scientific facts. However, to some this has led to the dogmatic pursuit of method over science and a development of disciplines at the expense of research capability (Feyerabend 1988, 2nd). This particularly problematic in cases where theories are considered incommensurable (Kuhn 1996, 3rd).

Within all of this disagreement about science, there is an acceptance of the importance of hypotheses when making objective inferences. Certainly, deriving theory from post hoc analysis, exploratory work or descriptive analysis is not considered good practice. This is not to say that in any study the framework for drawing conclusions should be completely specified before data collection begins; although this is considered appropriate practice when conducting randomised controlled trials (RCTs), it precludes the secondary analysis for which most large scale social surveys are designed. It would also put unnecessary constraints on longitudinal analyses, particularly allowing time into a study makes such rigid control infeasible. Feyerabend (1988, 2nd) has denigrated the emphasis on method as opposed to

pragmatism and Goldthorpe (2007, 2nd) appeals for more critical assessment of quality: all are agreed that science cannot be undertaken lightly.

There are established methods for the quantitative analysis of sequence data (see §3 for description of the nature of sequence data), such as complex (auto-) regressions, event histories and Markovian processes. These are able to test hypotheses with some statistical validity but such hypotheses are limited to transitions or short sequences as the amount of data required to identify a model increases exponentially with the length of the sequences involved (see §5). Therefore sequence hypotheses have not been investigated, and insofar as data and availability of methods drive research (see §2 for further discussion with relation to sequences), they have not been considered in a scientific fashion. Although sequence analysis purports to study holistically, the hypotheses expressed are almost without exception those normal to simple stochastic methods. Thus sequence analysis has been simply exploratory, and rarely makes comparisons between frameworks for analysis of the same data (see §5 for further discussion of the various methods and frameworks for analysis and the comparison of their results).

Substantial criticism has been levelled at the whole analytical frameworks presented as sequence analysis (Bakeman, Robinson, and Quera 1996;Levine 2000, 29;Wu 2000). This has led to some reflection on the algorithms used but little consideration of data or inference, similar to the concern expressed by Goldthorpe about the conflict between qualitative and quantitative research methods (see §2). To criticise a descriptive approach because it is not inferential is fatuous but it is also strange to contrast its success with that of an established approach with a strong methodological literature. Specific concerns have been: that there are better methods, that work has been unsuccessful, and that inferences are intrinsically lacking in validity. Sequence analysis seeks to address theories which are excluded by assumption in standard stochastic approaches: it does not challenge the results of other methods but seeks to understand theories which are incommensurable with them. Hence any negative results about hypotheses would be confounded with the possibility of the failure of the approach being taken thus very clear expectations of the data and possible inferences are necessary to refine the method. There has been poor application, as in all research, especially when the methods are novel, some improvements are needed in the

methodology (see §5), and more appropriate hypotheses should be chosen to make inferences about sequences.

Sequences are new data types and are often analysed without substantial preparation but this is partly because sequence hypotheses are difficult to formulate (see §2). Furthermore, once specified they are difficult to make inferences about as there is little understanding of the power involved in any analysis and therefore the requirements of the data, exploring sequence data systematically and ordering subjects purposively has only just been developed as a tool in sequence analysis (see Piccarreta and Lior 2010, 173). However, this is not a reason to avoid these challenges as there is considerable interest from several areas in making life course analyses and validating typologies of people in populations (Levy, Ghisletta, Le Goff, Spini, and Widmer 2005). There are numerous theories which purport to consider a latent characteristic of behaviour which may make certain simple events more likely but are considered to be distinctive across the whole individual experience. This is why a holistic approach is used to examine the whole sequence of experience of individual subjects.

Social theory often considers typologies as a means of understanding variation between individuals in their behaviours. For example, marketing literature has quite sophisticated profiles of consumer types, and learning styles are an accepted topic in education. The more theoretical conception of these typologies in generality is the ideal type (Prandy 2002). This is the theoretically grounded concept derived from the stereotype and may often be contentious but is certainly valuable in stimulating debate. Ideal types therefore provide a basis for hypotheses on sequences albeit challenging to operationalise, which has found application (Wiggins et al. 2007). This conceptualisation also exemplifies the concern for simple ideas in order to communicate results with an audience, something the method in general is lacking (see §7), particularly as the arbitrary nature of clusters and typologies is questionable (see §5 for further discussion on this point).

Sequence analysis has some utility in data mining (Hay, Wets, and Vanhoof 2004), whereas many wish it to use its results to challenge established theories (Abbott 1995). Although there is not a substantial barrier to new ideas in science, particularly

social science, Kuhn (1996, 3rd) argues rather convincingly that a new theory which in some way challenges an existing one needs far stronger support, especially to win over those working within the relevant discipline. Thus, although the analysis of sequences is considered a valid approach for investigating patterns in data, particularly new data or those resistant to other methods (for example web logs have no competing mode of analysis for individual user behaviour (Pallis, Angelis, and Vakali 2007)), it is not accepted, particularly from an inferential point of view, where other methods have succeeded (Wu 2000, 29). Yet sequence analysis uses tools whose application complements other analyses, enhances the understanding of the data, and allows for other interpretations (Abbott 2000).

A specific theory that shows obvious potential for analysis with sequences rather than simple stochastic models is the dual concept of social structure and individual agential behaviour (Marshall 2005). Most study focuses on particular transitions and the related social policy or social convention, naturally driven by the individuals concerned about their transitions or others sharing their concern or responsible for them at some level or other. However, most academic theory comes from a rather wider perspective of the trajectory followed by an individual, such as their social mobility, the structures which exist to guide them in certain directions, their own agency to beat their own path, and the final impact this all has on the complete trajectory. An instructive example of this is an Optimal Matching of career trajectories of women in finance (Blair-Loy 1999, 104) which shows up an agential cluster more likely to reach the top of the profession but very unlikely to follow any normal ideas of career structure, neither working up through one firm nor switching firms to advance themselves.

The concept of agency does not readily lend itself to hypotheses, especially not those of a kind which might easily be tested statistically. Part of this problem is in moving from a theory to a hypothesis in a particular case and therefore encapsulating the hypothesised difference in collectible data. Indeed a principal barrier here is the complex issue of time, one which is frequently mentioned as being fundamentally important in sequence analysis (e.g. Abbott 2001;Wu 2000, 29) but often mismatched between data and hypothesis. For example, data are recorded at the point when transitions become irreversible like marriage or graduation, but social theory may be

much more interested in partnership or education over the life course, but it is the end of the transition which is marked by the change of status (see §3). A bigger problem is the difficulty of specifying, and then identifying in the data, variation defined by being different to the rest, but representations will be useful for this purpose when combined with clustering techniques (see §7).

Theory often considers the mechanisms at work and the relative significance of different delays between transitions. It may well also interpret the interrelation of multiple sequences (see Pollock 2007, 170) as rather less coincidental than the juxtaposition of rites of passage. Therefore it may be necessary to consider the timing of events rather more psychologically (see Poincaré 1913 for details) in order to assess a hypothesis or warp the timings between events if this seems appropriate (e. g. Abbott and Hrycak 1990), or even discard the timings completely if it is really the orderings of events that is important. Certainly there is little point in adhering to one particular method (Feyerabend 1988, 2nd), especially if this has not been shown to have utility in addressing the problem at hand (Goldthorpe 2007, 2nd). Sequences contain enormous amounts of data but the best approach is to make different assumptions to those used in other analyses, rather than pretending to use none at all.

This discussion may seem to carefully skirt the issue of how one might make an inference about any given hypothesis. However, this is simpler than would appear, the only problem being in obtaining suitable data: having this, one can compare the variation hypothesised with what is evident in the data and develop other possible effects needing further investigation, more focussed hypotheses (Meehl 1967), and possibly new theory to reflect the difference between prediction and outcome. Indeed this is how some observers would say that science proceeds (e.g. Kuhn 1996, 3rd) and how qualitative work does now (Goldthorpe 2007, 2nd). The words of Joseph Kruskal are particularly relevant here:

"in many examples, it is remarkable that the differences are so small … Such close similarity can only be maintained by complex and subtle mechanisms."

(Kruskal 1983:9)

The key is to establish a framework for the analysis which will identify both similarity and differences for whole sequences to give some variation about which to make inferences.

The problems which remain therefore are the framework with which to analyse the data (see §5), and the mode of presentation to see the results of this oneself and elucidate them to others (see §7). Most specifically we need a mode of presentation with the elucidation of the theory at its heart in order to make any inferences, and if we are not used to thinking of things sequentially we should work harder to assist ourselves. Sequence analysis may be a different way of thinking but the main reason little progress has been made is that linking up their thinking through sequences has not been done (see §2).

5. Analytical Frameworks

When making analyses of data scientifically, it is understood that there needs to be some kind of framework with reference to which this is done. This has several advantages: people might replicate the work; they may understand more readily for its coherence; it can be compared, and contrasted, with the framework used in other work; and it hence allows it to sit as a building block in scientific understanding, e.g. as a special case open to generalisation. This does not mean there is only one way of performing any particular research, simply an understanding of a need for independent validation of the quality of the work (Goldthorpe 2007, 2nd).

For our purposes, in considering sequence analysis, we characterise an analytical framework as being a list of assumptions which are not tested by the study. By this we mean they are common to all analyses in the particular study, that they characterise the family of analyses being performed. It will be quite familiar that there is a difference between qualitative and quantitative frameworks but our definition is more practical; working from the bottom up, rather than the top down: it considers only those methods being used at any time; it is reflexive rather than relative. In this way it uses the concept of commensurability or consensus as defined by philosophers of science (Feyerabend 1988, 2nd;Kuhn 1996, 3rd)

When analysing sequences, as defined in §3, there is a clear division between researchers who choose to use stochastic models and those using algorithmic differencing. The former are things such as event histories (Wu 2004) and Markovian models (Bartholomew 1973) which make an assumption that there is some kind of limit to the information available from each sequence important to the model: elements far apart in the sequence don't affect each other; either through the number of earlier states affecting the current one, or the independence from the future. The latter are exemplified by Optimal Matching (Abbott 2000, 29) which assumes that the difference between whole sequences can be estimated consistently and thus the similarity structure of the subjects investigated (see, for example Abbott 1995, 21 for details). This does not preclude applying both frameworks to some data in one study and transcending the choice between them, as Abbott (2000, 29) might suggest with his 'toolkit' approach; however, the focus of this section is the limitations of each framework and its particular utility in relation to the other four elements.

The focus of both frameworks derives from a desire to understand any particular set of sequences, that sequences track the evolution of an individual over time and that it would be valuable to comprehend that evolution (see §2). The issues at hand are the particular sequence of states followed by individuals and the amount of time spent in each episode (see §3), and the strength of the association between these two with past and future evolution of the subject and how this can be represented so as to be understood (see §7). There is also the issue of the aggregate structure of the evolution of the subjects with respect to some fixed latent or unobserved variables. Thus both frameworks wish to consider the inherent structure of the sequence population with a view to some hypothesis derived from some theory (see §4).

In attempting to fit stochastic models to data, one has a distinct advantage of the maturity of the framework and the numerous options available (Wu 2000, 29). This maturity extends to a knowledge of the relative assumptions of the models so that these can be contrasted and take some account of the theory. There is also an understanding that the model is not the true behaviour of the system but it is a reflection of the observed data encapsulated in a form which is readily interpreted: all inferences are derived from observed probabilities of changing states. In fact this interpretation is a key feature such that predictions or simulations can be made from

these models and these represented in a variety of ways (see §7 for discussion of some practical limitations). Their wide implementation and history means they present a ready tool for making inferences predicated on the aforementioned framework.

The formidable background is a weakness for the stochastic models. Having established something reliable and versatile (for various examples see Bartholomew (1973, 2nd)), it is applied to every problem possible with adjustments included even for typical shortcomings like incomplete data (normally non-response and attrition for sequences but period data may be doubly censored). However, these methods are data driven so some theory does not fit very well (see §2). Faced with the common concern that long sequences are not fitted very well, methods to include latent variables and improve identifiability have been developed although these remain unsatisfactory as they are difficult to estimate (Green, Lid Hjort, and Richardson 2003). Other attempts reducing to conditionally independent effects are equally unsatisfactory (see Raftery and Tavare 1994) and essentially equivalent to those of Bakeman (c.f. Bakeman and Gottman 1997). There is a need to make confident inferences, rather than identify exact causal links, indeed external validity is much more important in social science when similar conditions would not be expected or enforceable.

The algorithmic comparison of sequences comes more from the idea of quantifying a life history approach to subjects than actually competing with stochastic models (Abbott and Forrest 1986); sequences are either identical or they have some differences. If they are different, we wish to describe how different they are: are all the same states present in both, in the same order and for the same duration? Some algorithms have been devised which transform one sequence to another, subject to some costs, and thereby calculate the difference between them, minimising over possible transformations (e.g. Abbott 1995, 21;Dijkstra and Taris 1995;Elzinga 2003;Piccarreta and Billari 2007). Others have considered common sub-sequences (Abbott and Barman 1997) or the time spent in each state (Dutreuil, Thibault, and Dutreuil 2008), and even made comparisons between selected algorithms (Elzinga and Liefbroer 2007, 23) but the general approach is to construct a dissimilarity matrix for the subjects.

This issue of what to do with the dissimilarity matrix once evaluated is a more open problem. Simply performing a cluster analysis and choosing a useful number of clusters is a common approach but lacks inferential validity, even if the corresponding dendrogram is used as a justification. Clustering around ideal types and model based clustering (for examples see Martin, Schoon, and Ross 2008 on ideal types; and Oh and Raftery 2007 on model-based clustering) offer much more obvious opportunities for inference, although better sequence hypotheses, as discussed in §4, would still afford much stronger conclusions. Thus many studies applying these methods result in typologies of doubtful validity and which are difficult to interpret (web logs, as mentioned earlier, being an extreme example), particularly as they are rarely well represented visually (see §7) leaving the whole approach open to criticism (Levine 2000, 29).

The assumptions made in the algorithm are particular to each algorithm and much discussion has focussed on choice of algorithm. For example, Optimal Matching assumes that every instance of a state is identical and can be swapped at a fixed cost for that of another. This is at odds with the intuition that an element representing a whole episode is rather more significant than one representing one tenth of the length of an episode. A criticism of this assumption, enhancement to the algorithm and an assessment of the difference it makes has been prepared by Brendan Halpin (2010). Other algorithms make assumptions but the variation between them can be compared with differences between the variable set used for stochastic models (the variation of outcome using different algorithms is not substantial, see for example Elzinga and Liefbroer 2007, 23).

Complex criticism of both frameworks takes issue with assumptions they make in their treatment of time. Assessments that algorithms ignore the role of duration of episodes are inaccurate but their current lack of attention to using continuous measurement or unequal timings is a weakness (Wu 2000, 29). More of a problem is the stationarity assumption applying to the whole process (this cannot be assumed when subjects come from different cohorts, especially when the theory investigates differences between cohorts c.f. Lesnard 2006, 2006-01). However, although time dependence is possible in stochastic models it is rarely identifiable, so it is the hypothesis, as described in §4, which needs to pay due regard to this problem. An

ideal development might be the availability of a topological approach to transformation between sequences, as proposed by Kruskal and Liberman (1983), although Lesnard (2006, 2006-01) still has concerns about a period approach, which certainly mean an appropriate hypothesis needs careful consideration (see §4).

Arguing on relative weaknesses as above helps to identify avenues for possible future research but is less useful in establishing the true utility of the frameworks (see §2). For these, it is best to consider which sequences one is looking at and the particular hypotheses involved (as described in §§3-4). Data in long sequences may result in the most appropriate stochastic models being impossible to identify in their entirety (a problem in structural equation models, c.f. McArdle 2005); short sequences may not show sufficient variation for algorithms (see Clark, Deurloo, and Dieleman 2003, 40), or place too much reliance on the assumptions or arbitrary costs. Moreover, hypotheses about conditions pertaining to particular transitions will be much better assessed by stochastic models whereas those looking at lifelong latent traits should be far more evident through within group similarity of their sequences. As always, complex hypotheses which make assumptions about the things they examine (like endogeneity) may not be testable at all; the principle is simply that the framework should be chosen by considering the situation one wishes to understand. Indeed the data and theory may yield the opportunity to apply different frameworks independently to the same problem (see §6 for more details).

A specific example in which differencing is more useful than stochastic modelling, is an assessment of the manifestations of individual agency in certain structures. The hypothesis being that there are expected trajectories into which subjects should cluster but the interest is in how some individuals deviate from these expected pathways. A stochastic model might identify the common transitions and patterns of such a structure but would not easily represent the deviation from it (although semi-parametric attempts could be made to estimate this). Of course there is still a problem of detecting outliers in cluster analysis, so there would still need to be a good understanding of the expected sequences in order to identify deviants.

Cohort data is the most natural medium for the similarities (Lesnard 2006, 2006-01) and period hypotheses will be more readily addressed with a stochastic model. Both

of these approaches make large numbers of assumptions, but adjustments and interpretative approaches do exist for the stochastic models. The algorithmic methods have no means of identifying the position in the trajectory and thus make little headway. Improved algorithms which allowed time warping and indexing of time would be a very valuable addition to the tools available which are limited to making many assumptions currently. Thus there are severe constraints on the hypotheses treatable by sequence analysis of period data, but the definition of a cohort may be much more subtle than the group of individuals of the same age: they need only be undergoing the same experience.

Most sequence analysis is limited by the data available as required sample sizes appear to be large, and the amount of data on each subject should also be substantial in order to secure the necessary variation and subtlety of the theory. Thus pre-existing data will impose certain limits on the analysis possible, notwithstanding the facility to adjust the categorisation of the data. There is also the opportunity to make more than one analysis of the same data, albeit considering different hypotheses, possibly derived from the same theory. Indeed this may be the best approach in making comprehensive inferences about the evolution of any set of subjects (Robette and Thibault 2008).

6. Categorisation

Having discussed the nature of sequences earlier, it may not be obvious that there is a need to return to an aspect of the data used in sequence analysis, less still that it should be fundamental to all analysis. However, if it is established that sequences, hypotheses and analyses are interdependent then an aspect of the data cannot be considered extrinsically fixed. That particular aspect is the categorisation of the state variable which is used to form the individual sequences. This is clearly distinct from and component to the sequences themselves but the categorisation is also dependent on the subject under study. It may also have an impact on the mode of analysis employed, especially in the case in which secondary data is being utilised. Thus the categorisation gives rise to important principles in any given inference. Indeed, it is the purpose of this section to expound the interplay between categorisation and the other elements of sequence analysis and thereby justify its inclusion as the associated principles are essential. However, it should be observed that the process of

categorisation is challenging, and although sequence analysis may put certain demands on the categorisation, its familiarity to the eventual audience will make a substantial difference in the dissemination of the results of the analysis (see §7). Many aspects of the process of categorisation are covered in detail by Bakeman (Bakeman and Gottman 1997, 2nd), and such issues as observation, validity and misclassification will only be discussed here in their relation to sequence analysis.

The primary issue in categorisation is encapsulated in what state variable to categorise in order to see variation between subjects, even when a categorical variable is already available from a secondary dataset. When to record a change of categories is also a concern, not least for the decision between stages, events and transitions (described in §3). The timing of the category should pay due regard to the hypothesised significance of the difference between categories, rather than the ease of measurement. Thus a simple and intuitive categorisation may be easy to implement but irrelevant or even obstructive the research being undertaken. One solution to this problem is to consider several state variables simultaneously (as in Pollock 2007, 170) which would capture the complexity of the theory but may need some reduction in order to ease the interpretation and visualisation of the result. Thus there will be interplay between the categorisation and representations (see §7).

Where hypotheses, as described in §4, seek to establish differences between subjects, the categorisation will need to be fine enough to exhibit these differences. Measurements will also need to be recorded regularly enough to make sure theoretically important episodes are not missed. That is, the subtlety and brevity of hypothetically important events should be accounted for in the data collection design. Indeed this will preclude the use of certain datasets to study certain phenomena, even though they may seem superficially well suited to such a study. However, it may be relevant ex-post to use the data in a coarser form than that originally collected, either through collapsing categories or some state combinations in multivariable approaches (Piccarreta and Lior 2010, 173). Social phenomena are complex enough that it is difficult to discard any data on the grounds that it is not relevant (c.f. Wu 2000, 29), so some ruthlessness will be necessary, but also several different analyses may be the best way of treating multiple hypotheses, rather than trying to draw many conclusions from a single analysis.

All categories should be substantively valuable, indeed it may be advantageous to have categories that overlap slightly, rather than having an exhaustive but uninterpretable category of 'other' (see Abbott and Barman 1997, 27 for example of the problems such a category can cause). If we consider categories as being some demarcation of a space, it should be clear that some categories will border each other, and thus be treated as more similar in the eventual model, whereas a category around and between the others will violate modelling assumptions of homogeneity. In this way a visual representation of the state space will be very useful (see §7).

A more subtle requirement of the categorisation is that it should be consistent throughout the dataset, or homogeneous and also invariant under the hypothesis. An assumption implicit in categorisation is that any instance of a category is in some sense equivalent to or exchangeable with any other. Of course it is unlikely to be necessary that they are all identical, although this would be ideal, a lack of equivalence may interfere with consideration of the hypothesis and any associated inference. For example, in comparing between two cohorts (as in Martin, Schoon, and Ross 2008) one may find that a difference in sequences observed may result from a change in the nature of instances of categories, rather than changes to the sequences themselves. Thus the hypothesis must be constructed with regard to the assumptions made by the categorisation and any systematic differences in categories.

The consistency of categories, mentioned above, is a more or less explicit assumption in any kind of analysis of the resulting sequences. In a stochastic process, for example, the probability of a transition from state **j** to state **i** is estimated based on the observed transitions, thus assuming these are in some way equal. In particular, it is normal to consider stochastic processes as being stationary i.e. independent of the time at which this transition took place (although §5 describes some embellishments to such a model, it also notes that many more data are required for useful estimation of time dependence). Similarly, algorithms normally make no reference to where in a sequence an element occurs, or at what time relative to others the sequence was experienced, when they calculate the difference between two sequences. These shortcomings reflect the genesic process described in §2 and are therefore avenues of current research, but also the fact that inferences can be made despite violated

assumptions (see Hwang and Green 2004 for evidence of the failure of this assumption in DNA mutation analysis). As for stochastic processes, the time invariance is not absolutely necessary but is often reasonable and will require careful data preparation to treat hypotheses where it cannot be assumed.

It was discussed in §5 that more than one analytical framework might be applied to the study of the same problem, albeit with different hypotheses. A key factor in this flexibility of application is the categorisation used: most simply, a finer categorisation will suit analysis by algorithmic differencing, where substantial variation between subjects is required (to avoid the problems experienced by Clark, Deurloo, and Dieleman 2003, 40); whereas coarser categorisations will suit stochastic models with hypotheses about specific transitions and limited numbers of parameters.

The life course approach is gaining increasing use in epidemiological study (Kuh and Ben-Shlomo 2004). This is not only due to understanding of the fetal origins of adult disease but also the various phases and stages that conditions and treatments go through. As an example, consider patients trying to give up smoking. This is a process which takes a considerable amount of time, and may involve periods of reduction in smoking, cessation, and relapse. This may be supported by various different treatments and substitutes, as well as being associated with an outcome. Developing a categorisation from this may seem reasonable enough, but there is always the problem of the level of smoking when smoking, as well as the exposure to other smoke-filled environments. More than this, though, there is the problem of the hypothesis: it is understood that some people succeed easily and others fail completely; the question is more related to the prognosis of those in between and effectiveness of interventions for them. Thus if the study wants to understand the interventions, the categorisation needs to be sensitive to the expected action of the intervention.

The most important point I argue about categorisation is that it is not fixed, and where it is almost determined by the availability of data this will impose limitations on the other elements of the analysis. However, we must realise that processes operate at several levels (Chomsky 1957), requiring different analyses and varied conceptual approaches to categorisation. In the analysis of web logs, we are presented with naturally categorical data, the individual pages visited, but these are a consequence of

the nature of the system and may not correspond usefully with the aspect of the user experience we wish to study (Iske 2009). Simply the categorisation should show sufficient variation between the subjects to make an inference about our hypothesis while allowing assumptions of invariance within categories to be excluded from our concern. The familiar and natural interpretability of the categorisation will remain important, particularly in the visual presentation and shall be discussed in the next section.

7. Representation

The aim of research is to advance understanding in such a way as to have some influence over practice. Thus, except in the case where the possible practitioners are all contained within the group of researchers (e.g. in an intelligence or otherwise confidential setting), there is a need to communicate the research undertaken. This communication may be to an audience already familiar with aspects of the work, and indeed already determined of the flaws in the work in hand. Hence a direct argument, although entirely reasonable and correct, may not exert the desired effect (Kuhn 1996, 3rd). Thus in this section we need to consider the process of inference for the individual, whether or not they are actually involved in the research, and the implications of this for the representation of sequence analyses.

It is more effective to present influential evidence in a mode considered objective (Tufte 1997) and do so openly, than to obscure or withhold the details. As §2 discussed, the path taken by studies in sequence analysis has been simultaneously to address research questions and promulgate the tools available. Despite persistence in these regards, there remain doubts as to the value of the research done (Levine 2000, 29), even though it is undeniable that they try to address interesting questions. Thus although the paper draws on some weaknesses of the preceding work to recommend an improved framework, it is obvious that some attention must be given to the communication of sequence analysis.

It is conventional to illustrate analyses with figures serving the purpose of complementing the explanation of the text, giving opportunity to different perspective and different ways of thinking. These figures can take the form of photographs, plots, diagrams and drawings and may feature any part of the research process from

conceptualisation and design to summaries of data or analyses. They are best used to express the unfamiliar in a more detailed and accessible form, and to show the conclusions in an evident fashion, exhibiting a visual explanation (Tufte 1997). Most quantitative analyses concentrate on the latter, due to the general familiarity of the problems but their conclusions are often complex and better suited to the naturally multivariate presentation of a chart or table.

Earlier sections have considered the necessity of understanding the nature of sequences and hypotheses involved. Thus we turn our attention away from the explicit conceptualisation and the communication implicit in that. What remains is a more intuitive understanding of sequence analysis: what a sequence looks like; what the results look like; what the results tell us about the hypothesis. To get these ideas across without forcing our view upon the reader requires a presentation both simple and detailed, rich in information, yielding insight without obscuring (Tufte 1997). One needs to be able to assess the difference between resemblance, contiguity and causality (Hume 2000:16), to allow us to see what the sequences are, how they are similar and how they are different. This is particularly important where the concepts involved are unfamiliar as we proceed with our previous experience in mind (Hume 2000). The representation should be clear enough for the viewer to see all of the detail while not giving a misleading impression of resemblance to other kinds of presentation

Sequence analysis is certainly not a familiar technique to substantive researchers due to its interdisciplinary applications and relative immaturity (see §2). This has led to a lack of standard visual representation, even in the reporting of stochastic models (although there are now some for OM in STATA (Brzinsky-Fay, Kohler, and Luniak 2006, 6)). Although thinking about sequences may not be a revolution of world view, it is certainly very different to the common concerns of quantitative analysis (Abbott 2000, 29) and may well have such an aspect when viewed up close (c.f. Kuhn 1996, 3rd). In this way it needs more than frames for the presentation of results but also description of the data, hypothesis and concepts involved; it also provides new tools for the researcher (Piccarreta and Lior 2010, 173). Visualisation will also aid the researcher, or would be researcher unfamiliar to the data or analysis, allowing initial exploration and stimulating curiosity. More open presentation would very likely avoid

the problems of arguments about facts (c.f. Hume 2000:24) which have been a major obstacle like that noted by Goldthorpe (Levine 2000, 29).

Huge tables of summary statistics have been produced in some reports (Bartholomew 1973, 2nd;Brzinsky-Fay 2007;e.g. Stark and Vedres 2006) and also detailed plots of events at various time lags, ostensibly showing significance (Altman 1965, 8). Most of these are lacking in a focus on either the sequences or hypotheses involved in the work, and although they might be valuable in a report following an established method, they lack the simplicity and complementarity to support the explanation in the text. Diagnostic plots have been produced, like the dendrogram from a cluster analysis (Everitt, Landau, and Leese 2001), and just as it should , this has aided those questioning the use of aspects of the algorithmic differencing methods (the main concern is that the distance displayed in a dendrogram has no interpretation c.f. Wu 2000, 29). It might be suggested that a more complex approach to clustering (Oh and Raftery 2001;Oh and Raftery 2007, 16) or multidimensional scaling (c.f. Piccarreta and Lior 2010, 173) are the best responses to this, which would also yield a presentation in more dimensions as well as allowing the examination of cluster density and relative positions of clusters.

Tufte (1997) has a rather high expectation of his viewers, demanding that "graphics should be as intelligent and sophisticated as the accompanying text", which does not make an allowance for a new and unfamiliar mode of analysis being applied to a familiar problem (c.f. Abbott 1995, 21). However, graphics are inherently multivariate which makes them ideal for presenting sequence analyses, especially given the effort of a bespoke design (as should be clear from the uniquely applicable one produced by Clark, Deurloo, and Dieleman 2003, 40). It must be cautioned that this complexity will make visual representations potentially misleading as they are identifying coincidental relationships (also a problem in cluster analysis, see Everitt, Landau, and Leese 2001, 4th). Indeed visual representations are inherently holistic and may flatter such an interpretation if the only presentations are comprehensive and lack focus on transitions.

More neglected than the presentation of results has been the use of representations as part of the research process in the algorithmic differencing approaches. Geographers

have of course produced some variety here (Gren 2001) but no link has been made to social science. These plots include representation of sequence spaces which are very difficult to read but have existed in geographical thinking since Hagerstrand (Gren 2001). The representation of the multivariate aspects of the data has recently been approached by Piccarreta (2010, 173), this even allows for assessment of the utility of categorisation, a particular problem experienced by Pollock (2007, 170) and likely to be useful elsewhere (Iske 2009). Such presentation would allow researchers also to demonstrate the validity of the choices they made in the analysis, with standard making modes for this, c.f. residual plots, making the methodology more open.

The issue of understanding the hypothesis and thus making inferences is far more complex than anything else in sequence analysis. Indeed, in cases of a study of structure, it is very difficult to conceptualise what agency would look like in the results, if it were to be there. This problem is present in the understanding of the sequences themselves and thus exploratory analysis of the data also. This is the point at which the choice of sequences, hypotheses and analytical frameworks may also become dependent on the capacity of associated representation to facilitate inferences. If inferences cannot be seen in the results of the analysis, the analysis itself is largely redundant, from an inferential point of view.

We can conclude that although representations are important, they should not be chosen to flatter the analytical framework preferred. Ideally, as discussed in §5, one might present figures from both approaches to analysis one might present figures from both approaches to analysis, for the contrasting conclusions evident therein but they may be better treated as incompatible but complementary in the understanding of the problem (as the analytical frameworks perform different roles in relation to the hypotheses, as described in §§3-5). Indeed, given the substantial conceptual differences between opposing frameworks, presentations may be the preferred medium for communicating their contributions in an open manner.

Our second conclusion would be about the need to display some accessible summary of the data. Several different approaches exist for this: multiple event histories can focus on stages (Brzinsky-Fay, Kohler, and Luniak 2006, 6) or events (Johnson 2004) but these are both cluttered and sometimes unnatural depending on the homogeneity

of the observations. Others try to show typical (Stark and Vedres 2006, 111) or ideal (Wiggins, Erzberger, Hyde, Higgs, and Blane 2007, 10) sequences; while others still are able to show a complete summary (Clark, Deurloo, and Dieleman 2003, 40) or summarise the generation process (Shoval and Isaacson 2007). The display of data will give some impression of the propriety of the categorisation for the needs of the analytical framework (the differences are discussed in §6) whereas there are some techniques for studying the frequency of short sequences and even corresponding exact tests (see Abbott and Barman 1997, 27;Bakeman, Robinson, and Quera 1996, 1 respectively). In all of this, long sequences remain very difficult to explore without an a priori construction (c.f. §4), indeed that is the main contention of Abbott (1995, 21).

Finally there is the problem of publication biases: results rarely reach those who might appreciate them. There is still the problem of disciplinarity (see §2) and the adherence to one particular analytical framework over solving the problems (see §5). Some work such as that on QHA has been published almost solely in French (Robette and Thibault 2008, 63). There is very little obvious place for methodological work on sequence analysis which has been predominately focussed on optimal matching, despite the range of other ideas (see Levy, Ghisletta, Le Goff, Spini, and Widmer 2005).

8. Principles

The preceding sections have argued that analysis of sequences is a relatively new and developing quantitative methodology, and as such makes rather different assumptions to established, cross-sectional approaches. Many of the studies published have been doubtfully received and some of the methodology used has received some criticism (Elzinga and Liefbroer 2007, 23;Hollister 2009, 38;Wu 2000, 29) including some attempts to improve the methodology used (e.g. Halpin 2010). Many of the concerns raised are legitimate, but their treatment has been rather piecemeal and failed to make a comprehensive analysis of the assumptions being made (c.f. Goldthorpe 2007, 2nd). Furthermore, the relative importance and interdependence of the assumptions has been ignored. Thus some concerns are essentially subordinate to others, including some which might in other analyses be considered fundamental, such as misclassification being a function of categorisation. So we are left with five elements

which encompass the key assumptions and principles which represent their interdependencies, schematically represented in Figure 1.

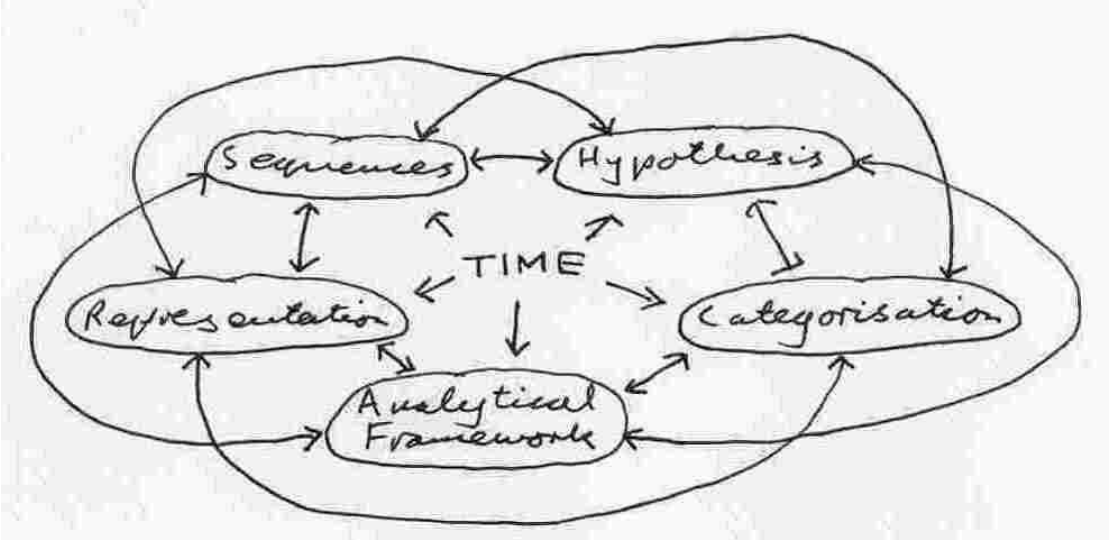

Figure 1: Interrelations between the five elements, with time at the centre

The sequences collected for study will be dependent on the hypothesis we intend to consider, like whether it involves transitions, events or stages. The availability of sequences will limit the hypotheses admitting to analysis, whether whole cohorts are available, or censoring or simple periods are available. Sequences will determine which analytical framework is identifiable in that sequences with small numbers of states and episodes will require stochastic models, and long ones may need other approaches. A decision to use a particular framework, for example to exhibit a new development, will determine requirements of the sequences to give an estimable number of parameters and sufficient variation between subjects. The assumptions of the framework will preclude the study of certain hypotheses thus the hypothesis needs to be chosen with regard to the intended inference. Similarly, a determination to use a particular framework will put constraints on the hypotheses which can be usefully considered and again the variation needed to test them. Thus the first three elements have a mutual interdependence which is obviously fundamental to any research process.

Categorisation of the state variable is obviously necessary to constructing sequence data at all. However, there is considerable flexibility open to the researcher in determining this, and this can be very important in the research process (Piccarreta

and Lior 2010, 173). The categorisation determines the sequence of each individual, like whether just one characteristic or several is used as the state variable. The hypothesis will require the categorisation to show the hypothesised difference between the individuals. It is the similarity between some individuals which is important (Kruskal 1983), but we must be able to tell the groups apart in the way hypothesised. Categorisation makes an assumption of homogeneity within categories and invariance over time which the analytical framework and the hypothesis must admit. The analytical framework will be dependent on the categorisation to give the appropriate level of detail to make any inferences and allow the relevant model to be identified. Thus the categorisation of the state variable is of considerable importance to the researcher but this power is contingent on sensitivity to the context of the research which may not indicate the use of a conventional categorisation, despite the benefits of its familiarity.

A balance must be found between analytical power of the bespoke and the explicative power of the familiar. Whereas for descriptive work it may be beneficial to make things simple for the analyst, that concern cannot be given pre-eminence in an inferential framework. Representations are fundamental to processes of analysis, whether this is the theory, the raw data, or the model summary which is being represented. Whenever choices are made, these should be represented clearly so that a decision can be made based on the evidence, rather than a predetermined idea, this also gives the result of a clear process for scrutiny of others. Thus all of the other elements are dependent on the use of representation, from the fundamental understanding of what the sequences are, so important in an emerging area, to the explication of a hypothesis, e.g. using ideal types, and choices of categorisation. While representation might seem part of the analytical framework, its importance affecting both all the other elements, and the presentation of work to an audience, means the subtlety of choices and development of such presentation is essential to any inference. New methods may be powerful, but without powerful representation their insight will be lost so some compromises will be necessary.

Thus the principles do not make for a complete interdependence but more a restraint on the blind analysis of data. The process of sequence analysis is inherently holistic on subjects' experience, and so considerations of the complex interrelation of the

elements should be fairly intuitive to practitioners. The principal difficulty in applying these principles is the lack of many successful inferential studies to draw ideas from. However, they should be used iteratively, and many studies give an obvious starting point from which to work, an element which is fixed by design. There is also a need to develop the tools available further, so that fewer compromises will be necessary.

9. Progress

Most success in the application of sequence analysis so far has been descriptive of the data analysed and has documented familiar kinds of lives and experiences. This paper attempts to establish an inferential framework but application of the principles should establish greater descriptive capacity in new and emerging areas of application. In fields where there is a tradition of discussing the states which make up sequences, like life course theory (Levy 2005), there is already a descriptive basis for the analysis of sequences. In the analysis of things like web logs, which is much less advanced, there has been very little success in even describing the trajectories followed by individuals or the similarities, differences and clustering patterns. Here there need to be established sequences which embody familiar ideas so as to develop resemblances to them in future work.

The principles make it clear that there are inferential weaknesses to existing methodology and previous many studies. Hence although sequence analysis has been criticised for its lack of contribution, it may be that it has yet to be developed sufficiently. Particularly the two analytical frameworks have weaknesses: the stochastic are too simple and the assumptions are very restrictive; the holistic algorithmic methods are less well established and their assumptions are not well enough understood to yield inferential framework. Both methods have to a large degree been poorly executed in the literature although we do not wish to propose a standardised method, indeed which would be at odds with the principles above. However, criticism of particular methods relates more to their clear explication of their approach, rather than low quality of the studies. Other problems like misclassification affect both frameworks and were discussed in §6, visual aspects of description are an apparent weakness of most sequence analyses (see §7), but there may be forgotten or neglected tools in other disciplines. One of the greatest advances

in both inferences and dissemination would be if there were more appropriate representations used in research and reported in publications.

In any field of research there can be problems when the users are uncritical of their methodology (Feyerabend 1988, 2nd). The principles espoused above may not be a solution to the problem of inference, but they surely aid understanding of whether the next steps are to develop methodology further, focus on different questions, or collect new data. Following principles of any kind is much more open to the understanding of others, whether or not they choose to agree with the methods followed. A more systematic, but not prescriptive (c.f. Feyerabend 1988, 2nd), approach will be of considerable help in addressing the conflict identified elsewhere by Goldthorpe (Goldthorpe 2007, 2nd). Even if there are substantial weaknesses to the principles, we would hope, as Chomsky did, that:

"By pushing a precise but inadequate formulation to an unacceptable conclusion, we can often expose the exact source of the inadequacy and, consequently, gain a deeper understanding of the … data."

(Chomsky 1957:5)

Overall it is best to be guided by a research question but careful of the assumptions made and the interplay between the two, given the constraints imposed by the availability and form of the data. Thus any sequence analysis should be developed with regard to principles and mindful that the resulting inference may be a better understanding of the problem, rather than a solution to it. Algorithms and models would then continue to develop as they have done: in conjunction with the problems they seek to understand. An engagement with the success that has happened and a truly holistic rather than disciplinary approach to problems, using both frameworks in the same analysis and looking far back to studies in things like life course epidemiology (Kuh and Ben-Shlomo 2004) may not be a comfortable idea to some, but it is likely to be the most successful.

References

Abbott, A. 1995. Sequence Analysis: new methods for old ideas. *Annual Review of Sociology* 21: 95-113

-----. 2000. Reply to Levine and Wu. *Sociological Methods and Research* 29 (1): 65-76

-----. 2001. *Time Matters: on theory and method*. Chicago: University of Chicago Press

Abbott, A, and E Barman. 1997. Sequence Comparison via Alignment and Gibbs Sampling: a formal analysis of the emergence of the modern sociological article. *Sociological Methodology* 27: 47-87

Abbott, A, and J Forrest. 1986. The Optimal Matching Method for Anthropological Data: An Introduction and Reliability Analysis. *Journal of Quantitative Anthropology* 2: 151-170

Abbott, A, and A Hrycak. 1990. Measuring Resemblance in Sequence Data: an optimal matching analysis of musicians' careers. *American Journal of Sociology* 96 (1): 144-185

Altman, SA. 1965. Sociobiology of Rhesus Monkeys: II stochastics of social communication. *Journal of Theoretical Biology* 8: 490-522

Bakeman, R and JM Gottman. 1997. *Observing Interaction: an introduction to sequential analysis*. Cambridge: CUP

Bakeman, R, and V Quera. 1995. Loglinear Approaches to Lag-sequential Analysis when Consecutive Codes may, and cannot, Repeat. *Psychological Bulletin* 118 (2): 272-284

Bakeman, R, BF Robinson, and V Quera. 1996. Testing Sequential Association: estimating exact p-values using sampled permutations. *Psychological Methods* 1 (1): 4-15

Bartholomew, DJ. 1973. *Stochastic Models for Social Processes*. London: Wiley

Billari, FC, and R Piccarreta. 2005. Analysing Demographic Life Courses through Sequence Analysis. *Mathematical Population Studies* 12 (2): 81-106

Blair-Loy, M. 1999. Career Patterns of Executive Women in Finance: an optimal matching analysis. *American Journal of Sociology* 104 (5): 1346-1397

Brzinsky-Fay, C. 2007. Lost in Transition? Labour market entry sequences of school leavers in Europe. *European Sociological Review*

Brzinsky-Fay, C, U Kohler, and M Luniak. 2006. Sequence Analysis with STATA. *The Stata Journal* 6 (4): 435-460

Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton

Clark, WAV, MC Deurloo, and FM Dieleman. 2003. Housing Careers in the United States, 1968-93: modelling the sequencing of housing states. *Urban Studies* 40 (1): 143-160

Cox, TF and MAA Cox. 2001. *Multidimensional Scaling*. Boca Raton: Chapman and Hall/CRC

Dijkstra, W, and T Taris. 1995. Measuring the Agreement Between Sequences. *Sociological Methods and Research* 24 (2): 214-231

Dutreuil, R, N Thibault, and C Dutreuil. 2008. Comparing Qualitative Harmonic Analysis and Optimal Matching: An Exploratory Study of Occupational Trajectories. *Population* 63 (4): 533-556

Elzinga, C. 2003. Sequence Similarity: a non-aligning technique. *Sociological Methods and Research* 32 (1): 3-29

Elzinga, C, and AC Liefbroer. 2007. De-standardization of Family-Life Trajectories of Young Adults: A Cross-National Comparison Using Sequence Analysis. *European Journal of Population* 23: 225-250

Everitt, BS, S Landau, and M Leese. 2001. *Cluster Analysis*. London: Hodder Arnold

Feyerabend, PK. 1988. *Against Method*. London: Verso

Goldthorpe, JH. 2007. Current Issues in Contemporary Macrosociology. In *On Sociology*, Stanford, CA: Stanford University Press

Green, PJ, N Lid Hjort, and S Richardson. 2003. *Highly Structured Stochastic Systems*. Oxford: OUP

Gren, M. 2001. Time Geography Matters. In *Timespace: geographies of temporality*, eds J May and N Thrift London: Routledge

Halpin, B. 2010. Optimal Matching Analysis and Life Course Data: the importance of duration. *Sociological Methods and Research*

Hay, B., G. Wets, and K. Vanhoof. 2004. Mining navigation patterns using a sequence alignment method.

Hollister, M. 2009. Is Optimal Matching Suboptimal? *Sociological Methods and Research* 38 (2): 235-264

Hume, D. 2000. *An Enquiry Concerning Human Understanding*. Oxford: Clarendon Press

Hwang, DG, and P Green. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Science* 101 (39): 13994-14001

Iske, S. 2009. Educational Research Online: E-Learning Sequences analyzed by means of Optimal-Matching.

Johnson, S. 2004. Event Chart Visualisation of NHS Direct Onling User Weblog Data: developing a methodology. MSc University of Southampton.

Kruskal, JB. 1983. An Overview of Sequence Comparison. In *Time Warps, String Edits and Macromolecules: the theory and practice of sequence comparison*, eds D Sankoff and J Kruskal, 1-44. Reading, MA: Addison-Wesley

Kruskal, JB, and M Liberman. 1983. The Symmetric Time-Warping Problem: from continuous to discrete. In *Time Warps, String Edits and Macromolecules: the theory and practice of sequence comparison*, eds D Sankoff and J Kruskal, 125-162. Reading, MA: Addison-Wesley

Kuh, D and Y Ben-Shlomo. 2004. *A Life Course Approach to Chronic Disease Epidemiology*. New York: OUP

Kuhn, TS. 1996. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press

Lesnard, L. 2006. Optimal Matching and Social Sciences. Paris: Crest, Insee.

Levine, JH. 2000. What have you done for us lately? *Sociological Methods and Research* 29 (1): 34-40

Levy, R. 2005. Why Look at Life Courses in an Interdisciplinary Perspective? In *Towards an Interdisciplinary Perspective on the Life Course*, eds R Levy, P Ghisletta, J-M Le Goff, D Spini and E Widmer, 3-33. Amsterdam: Elsevier

Levy, R, P Ghisletta, J-M Le Goff, D Spini, and E Widmer. 2005. *Towards an Interdisciplinary Perspective on the Life Course*. Amsterdam: Elsevier

Marshall, VW. 2005. Agency, Events, and Structure at the End of the Life Course. In *Towards an Interdisciplinary Perspective on the Life Course*, eds R Levy, P Ghisletta, J-M Le Goff, D Spini and E Widmer, 57-92. Oxford, UK: Elsevier

Martin, P, I Schoon, and A Ross. 2008. Beyond Transitions: applying optimal matching analysis to life course research. *International Journal of Social Research Methodology*

McArdle, JJ. 2005. Five Steps in LAtent Curve Modeling with Longitudinal Life-Span Data. In *Towards an Interdisciplinary Perspective on the Life Course*, eds R Levy, P Ghisletta, J-M Le Goff, D Spini and E Widmer, 315-360. Oxford, UK: Elsevier

Meehl, PE. 1967. Theory-testing in psychology and physics: a methodological paradox. *Philosophy of Science* 34: 103-115

Oh, M-S, and AE Raftery. 2001. Bayesian Multidimensional Scaling and Choice of Dimension. *Journal of the American Statistical Association* 96 (455): 1031-1043

-----. 2007. Model-based Clustering with Dissimilarities: a Bayesian approach. *Journal of Computational and Graphical Statistics* 16 (3): 559-585

Pallis, G, L Angelis, and A Vakali. 2007. Validation and Interpretation of Web Users' Sessions Clusters. *Information Processing and Management* 43: 1348-1367

Piccarreta, R, and FC Billari. 2007. Clustering Work and Family Trajectories by using a Divisive Algorithm. *Journal of the Royal Statistical Society A* 170 (4): 1061-1078

Piccarreta, R, and O Lior. 2010. Exploring Sequences: a graphical tool based on multi-dimensional scaling. *Journal of the Royal Statistical Society A* 173 (1): 165-184

Poincaré, JH. 1913. *The Value of Science*. New York: Dover

Pollock, G. 2007. Holistic Trajectories: a study of combined employment, housing and family careers by using multiple sequence analysis. *Journal of the Royal Statistical Society A* 170 (1): 167-183

Prandy, K. 2002. Ideal Types, Stereotypes and Classes. *British Journal of Sociology* 53 (4): 583-601

Raftery, AE, and S Tavare. 1994. Estimation and Modelling Repeated Patterns in High Order Markov Chains with the Mixture Transition Distribution Model. *Applied Statistics* 43 (1): 179-199

Robette, R, and N Thibault. 2008. Comparing Qualitative Harmonic Analysis and Optimal Matching: An Exploratory Study of Occupational Trajectories. *Population* 63 (4): 533-556

Sankoff, D and JB Kruskal. 1983. *Time Warps, String Edits and Macromolecules: the theory and practice of sequence comparison*. Reading MA: Addison-Wesley

Shoval, N, and M Isaacson. 2007. Sequence Alignment as a method for Human Activity Analysis in Space and Time. *Annals of the Association of American Geographers* 97 (2): 282-297

Stark, D, and B Vedres. 2006. Social Times of Network Spaces: network sequences and foreign investment in Hungary. *American Journal of Sociology* 111 (5): 1367-1411

Tufte, ER. 1997. *Visual Explanations: images and quantities, evidence and narrative*. Cheshrire, CT: Graphics Press

Wiggins, RD, C Erzberger, M Hyde, P Higgs, and D Blane. 2007. Optimal Matching Analysis Using Ideal Types to Describe the Lifecourse. *International Journal of Social Research Methodology* 10 (4): 259-278

Wilson, WC. 1998. Activity Pattern Analysis by Means of Sequence-Alignment Methods. *Environment and Planning A* 30 (6): 1017-1038

-----. 2006. Reliability of Sequence-Alignment Analysis of Social Processes: Monte Carlo tests of ClustalG software. *Environment and Planning A* 38: 187-204

Wu, LL. 2000. Some comments on 'Sequence analysis and optimal matching techniques in sociology: review and prospect'. *Sociological Methods and Research* 29 (1): 41-64

-----. 2004. Event History Models for Life Course Analysis. In *Handbook of the Life Course*, eds J Mortimer and M Shanahan, 477-502. New York: Springer