Statistical and Stochastic Analysis of Sequence Data

Ming Ming Chiu and Peter Reimann

Abstract Two common CSCL questions regarding analyses of temporal data, such as event sequences, are: (i) What variables are related to event attributes? and (ii) what is the process (or what are the processes) that generated the events? The first question is best answered with statistical methods, the second with stochastic or deterministic process modeling methods. This chapter provides an overview of statistical and stochastic methods of direct relevance to CSCL research. Many of the statistical analyses are integrated into statistical discourse analysis. From the stochastic modeling repertoire, the basic hidden Markov model as well as recent extensions is introduced, ending with dynamic Bayesian models as the current best integration. Looking into the near future, we identify opportunities for a closer alignment of qualitative with quantitative methods for temporal analysis, afforded by developments such as automization of quantitative methods and advances in computational modeling.

Keywords Statistical discourse analysis · Time analysis · Stochastic models · Process mining

1 Definitions and Scope

In this chapter, we introduce two complementary approaches for the analysis of temporal data, in particular for the analysis of discrete event sequences: statistical and stochastic analysis. The basic distinction is that a stochastic process is what (one assumes) generates the data that statistics analyze. To say that a process is

M. M. Chiu (✉)
Special Education and Counseling, The Education University of Hong Kong, Tai Po, Hong Kong
e-mail: mingchiu@eduhk.hk

P. Reimann
Centre for Research on Learning and Innovation, University of Sydney, Sydney, Australia
e-mail: peter.reimann@sydney.edu.au

U. Cress et al. (eds.), *International Handbook of Computer-Supported Collaborative Learning*, Computer-Supported Collaborative Learning Series 19,
https://doi.org/10.1007/978-3-030-65291-3_29

"stochastic" is to say that at least part of it happens "randomly"; it can be studied using probability theory and/or statistics. The analysis of stochastic processes is the subject of probability theory, like statistics, a field of study in mathematics. In probability theory, we have some given probability distribution and want to determine the probability of some specific event.

The following sections will be introducing regression models as a powerful statistical modeling method and hidden Markov models as an example of a stochastic method. In their combination, they can be used for both empirical and theoretical modeling.

1.1 Statistical View of Sequential Processes

Computer-supported collaborative learning (CSCL) researchers often ask five types of questions that involve time: (a) are there common sequences of actions/events (e.g., *disagree* → *explain*)? (b) do these sequences have antecedents at various levels? (c) are there pivotal events? (d) do these sequences differ across time periods? and (e) are these sequences related to outcomes? First, are disagreements more likely than other utterances to be followed by explanations in an online forum? These types of questions ask whether one event (e.g., *disagree*) is more likely than otherwise to be followed by another event (e.g., *explain*, Chiu 2008).

Second, are factors at other levels (e.g., *gender* of author or recipient; *mean writing grade of group*) related to the likelihood of a *disagree* → *explain* sequence? Such questions help build a comprehensive theoretical model of the different attributes across levels that might influence the likelihood of such sequences (Chiu and Lehmann-Willenbrock 2016).

Third, does a pivotal action/process (e.g., *summary*) change the likelihood of a *disagreement* → *explanation* sequence across time (Wise and Chiu 2011)? Such questions seek to identify actions/events that radically change the interaction (*pivotal events,* Chiu and Lehmann-Willenbrock 2016).

Fourth, are *disagree* → *explain* sequences more likely at the beginning, middle, or end of a discussion? Such questions ask whether a particular sequence is more likely at different time periods, thereby examining their generality across time (Chiu 2008).

Lastly, do groups with more *disagree* → *explain* sequences than others show superior group solutions or subsequent individual test scores? Such questions help build a comprehensive theoretical model of the consequences of such sequences for groups/individuals (Chiu 2018).

Before proceeding further, we define several terms: sampling unit, session, time period, event, and sequence. The object under study (group, dyad, or individual) is the *sampling unit*, which is observable during one or more *sessions* (occasions). If warranted, we can divide each session into *time periods*. During a session, we observe one or more learners' behaviors, which we call *events*. One or more adjacent events is a *sequence.*

A statistical analysis of data that address the above research questions has three major assumptions (Teddlie and Tashakkori 2009), which we explicate in the

context of online students chatting about designing a paper airplane to stay aloft longer. First, instances of a category (e.g., *disagree*) with the same value (e.g., disagree vs. not disagree [coded as 1 vs. 0]) are sufficiently similar to viewed as equivalent. Second, earlier events (*disagree* in parent message 87) or fixed attributes (e.g., *author gender*) influence the likelihood of a specific event at a specific time (*explanation* in message 88). Third, our statistical model fully captures our theoretical model, so that unexplained aspects of the data (*residuals*) reflect attributes that are not related to our theoretical model.

1.2 The Stochastic View of Sequential Processes

The stochastic perspective of sequential data in CSCL assumes that a recorded sequence—for instance, a sequence of dialogue moves—is produced by a stochastic process. Events are seen as different in kind from processes: Processes produce (generate, bring about) events. While recorded events can be analyzed to identify structure and properties of processes, they are not identical with the latter. The ontological position that processes are different from events is foundational to stochastic (and deterministic) models, but it is not shared by regression models and most other variants of the general linear model, with the exception of structural equation models under a certain interpretation of what latent variables mean (Loehlin 2004). Regression models' variables are ontologically "flat"; the only difference between them is epistemic: the variation in the dependent variable is explained in terms of the covariation with one or more independent variables. Note that multilevel modeling (Cress 2008) does not change the ontological status of the variables included either: The *nesting* relation in multilevel modeling is different from the *generative* relation that links structure/process to events.

What are the "practical" consequences of this distinction for the learning researcher? For one, stochastic models are not dependent on distribution assumptions, such as normal distribution. Secondly, stochastic modeling allows to simulate the implications of changes to theoretical assumptions; they afford counterfactual ("what if?") reasoning. And thirdly, with this kind of model one can determine the likelihood of an individual event sequence being producible by the process the model describes. Thus, they are not so much an alternative to statistical models than they allow to answer additional questions.

2 History and Development

2.1 Early Statistical Analyses

Early researchers analyzed their data with simple, mathematics calculations, namely conditional probabilities. To test hypotheses, researchers developed statistical

Table 1 A comparison of conditional probabilities, sequential analysis, and regressions

Properties	Conditional probability	Sequential analysis	Vector auto-regression
Discrete outcomes (explain vs. not)	✓	✓	✓
Discrete explanatory variables	✓	✓	✓
Significance test		✓	✓
Goodness of fit		✓	✓
Continuous outcomes			✓
Continuous explanatory variables (notably *time*)			✓
Explanatory variables at other levels			✓
Nonconsecutive events			✓
Complex models			✓
Small sample size	✓		✓

methods, such as sequential analysis and regressions (see Table 1). We explicate these methods using examples from online students chatting about paper airplane design.

2.1.1 Conditional Probability

The probability of an event (e.g., explain) given that another event (disagree) has occurred is its *conditional probability* (CP, e.g., Farran and Son-Yarbrough 2001). To compute it, we divide the overall probability (OP) of the *disagree* → *explain* sequence (e.g., 13%) by the overall probability of disagreeing (e.g., 39%), yielding 33% (=13%/39% = OP [*disagree* → *explain*]/OP[*disagree*]) via *Bayes' theorem*. CPs apply to sequences of any length. However, CP has no significance tests or goodness-of-fit measures, so researchers must subjectively decide whether a CP supports or rejects their hypotheses.

2.1.2 Sequential Analysis

To test hypotheses, researchers developed statistical methods, such as sequential analysis (SA). Building on CP, SA supports hypothesis testing. SA models events across time as a discrete process in which the current event (*state*) determines the probability of the next event (Gottman and Roy 1990). For example, a group in a state of *disagreement without explanation* is more likely than otherwise to move to a state of *disagreement with explanation* (Chiu and Lehmann-Willenbrock 2016). SA tests for significant differences (*z-score*) and evaluates the goodness-of-fit of each explanatory model (via *likelihood ratio chi-squared* tests, Bakeman and Gottman 1986). Like CP, SA only applies to discrete outcomes and explanatory variables at

the same level (message), requires consecutive events, and can require enormous sample sizes to test somewhat complex explanatory models.

2.1.3 Vector Auto-Regression

Addressing these three limitations of CP and SA, vector auto-regressions (VAR, Kennedy 2008) model continuous variables, explanatory variables at different levels, nonconsecutive events, and complex phenomena with small samples. For a continuous outcome variable, an *ordinary least squares* regression fits a line to the data (or more generally a curve), which enables analyses of outcomes as a function of time (traditional time-series data, Kennedy 2008). For a dichotomous outcome (explanation vs. no explanation), a *logit, probit,* or *gompit* regression fits an S-curve to the data (see Fig. 1 for an example with a continuous, explanatory variable *age*; Cohen et al. 2003). This example also shows how regressions can test explanatory variables at any level (message, person, group, etc., Kennedy 2008). Furthermore, regressions can model nonconsecutive relations, such as whether a student who disagreed two messages ago (grandparent message: disagree [−2]) raises the likelihood of an explanation in the current message (explain), namely, disagree (−2) → → explain (Chiu and Lehmann-Willenbrock 2016). In general, we can test whether an attribute of an earlier event is related to an attribute of the current event (Kennedy 2008).

Also, a regression can create simpler models of complex phenomena via multidimensional coding. For example, to model sequences with five events from four dimensions with two choices per dimension (e.g., female [vs. male], student

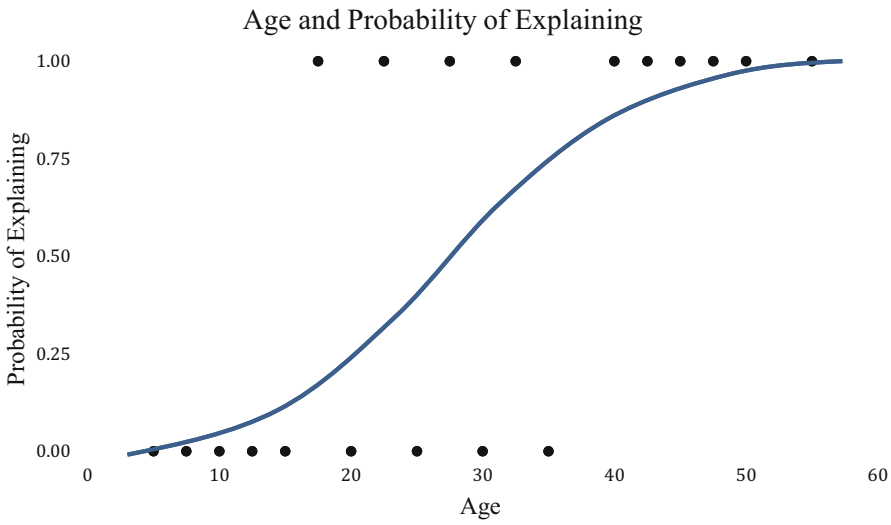

Fig. 1 Logit regression fitting an S-curve to data on age and explanation

[vs. teacher], disagree [vs. agree], and explain [vs. not], SA requires a sample size of 5,242,880 (=5 × [2^4]^5; = 5 × [*event combinations*]$^{sequence\ length}$, Gottman and Roy 1990). In contrast, a regression only requires 20 explanatory variables (20 = 4 dimensions × sequence length of 5) and a much smaller sample (Cohen et al. 2003). Applying Greene's (1997) sample size formula for regressions ($N > 8 \times [1 - R^2]/R^2 + M - 1$; with expected explained variance $R^2 = 0.1$ and number of explanatory variables $M = 20$), testing this explanatory model requires a sample size of only 91 (=8 × (1 − 0.10)/0.10 + 20 − 1). Hence, a multidimensional coding scheme can capture the complexity of the model, reduce the number of needed variables, and reduce the minimum sample size needed for a regression (Chiu and Lehmann-Willenbrock 2016).

2.2 Early Applications of Stochastic Analysis

An important method to modal temporal data probabilistically is the hidden Markov model (HMM). It has been applied in CSCL research for analyzing discourse sequences, for example. This formalism is an extension of the (discrete) Markov Process model, which we introduce first.

2.2.1 Markov Models

The underlying assumption of probabilistic models is that the event sequence can be characterized as a parametric random process and that the parameters of the stochastic process (the structure, not the event sequence) can be determined (estimated) in a precise, well-defined manner (Rabiner 1989, p. 255). A Markov process model describes a system made out of N distinct states, S_1, S_2, …, S_N. At equally spaced discreet times ($t = 1, 2, \ldots$), the system undergoes a change of state, with the state at time t denoted as q_t. A full description of the system would require the specification of the current state (time t) as well as all the predecessor states. For the important special case of a discrete first-order Markov chain, it is assumed that this description can be truncated to just the current and the predecessor state, i.e.,

$$\mathrm{P}\left[q_t = S_j\right] = \mathrm{P}\left[q_t = S_j | q_{t-1} = S_i\right].$$

We further assume that the transitions between states are independent of time, that the system itself doesn't change over time. This leads to a set of state transition probabilities a_{ij} of the form.

$$a_{ij} = \mathrm{P}\left[q_t = S_j | q_{t-1} = S_i\right], 1 \leq i,j \leq N$$

and with the property that the sum of all transitions probabilities across the states S_j is equal to 1:

$$\sum_{j=1}^{N} a_{ij} = 1$$

To provide an example, let's assume we want to describe a (hypothetical) group with three states: (1) forming, (2) storming, or (3) norming (Tuckman 1965). Recording observations of group communication as they unfold, we can describe the system in terms of transition probabilities between these three states:

$$A = \{a_{ij}\} = \begin{Bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{Bmatrix}$$

The value in middle, 0.6, for instance, means that the probability that if the group is in the storming phase at time t_i it will be in that phase at time t_{i+1} as well is 0.6; and the 0.2 to the left refers to the chance of changing from forming to storming. One question this model can be used to answer is: What is the probability of the group over the next days being "forming–storming–forming–norming. . .," or any other specific sequence? Another question that can be answered from the model is: Given that the model is in a known state, what is the probability that it will stay there for exactly t number of interactions? Note that the transition matrix is also the place where theoretical assumptions can be varied, to the extent that they can be expressed as (transition) probabilities. For instance, to express that it should not be possible to move from forming to norming directly one can set the corresponding transition probability to a very small number.

2.2.2 Hidden Markov Models

With HMMs, we can account for the *relation* between states and observed events by making the observation a probabilistic function of the state. The resulting hidden Markov model is ". . .a doubly embedded stochastic process with an underlying stochastic process that is not observable (it is hidden), but can only be observed through another set of stochastic processes that produce the sequence of observations" (Rabiner 1989, p. 259). To describe an HMM, we need to specify (1) the number of states in the model, (2) the number of distinct observation symbols per state, i.e., the observables; (3) the *state* transition probability distribution, (4) the *observation* probability distribution for each state, and (5) the initial state distribution (a vector with the probabilities of the system being in state j). Based on this specification, an HMM can be used in three main ways: (a) for generating ("predicting") observations; (b) for modeling how a given observation sequence

was generated by an appropriate HMM; (c) for parsing a given observation sequence and thereby deciding if the observed sequence is covered by (or explained by) the model.

The best-known early example of HMM use in CSCL is likely (Soller 2004). Here, chat contributions from pairs of learners involved in a problem-solving task were first coded into categories and HMM models were then trained on the chat sequences (observations) for successful and unsuccessful pairs, respectively. The method proved useful to identify sequences that led to successful knowledge sharing from those that did not. In general, training an HMM on known observations requires specification of the initial state probability distribution as well as the state transition matrix and observation probability distribution for each state. Based on these initial specifications, programs such as seqHMM (Helske and Helske 2017) can calculate values for parameters in the HMM that lead to a best fit with the observed data. Boyer et al. (2009) used HMMs in a similar fashion. Here the pairs were formed by a student and a tutor, and the dialogue acts were coded in terms of categories relevant for tutor-student discourse. States were interpreted as "dialogue modes." As in the Soller study, the number of states were determined by balancing the number of states with the fit to training data.

3 State of the Art

3.1 Recent Developments in Statistical Analysis

CP, SA, and simple regressions all assume: (a) sequences, (b) identical task difficulties, (c) no group/individual differences, (d) no time periods, (e) a single outcome, (f) observed events only, (g) direct effects only, (h) no measurement error and (i) an immediate outcome (see Table 2). Specifically, researchers addressed them via:

Table 2 Analytic issues and suitable statistical strategies

Analytic issue	Statistical strategy
Parallel chats or trees	Store parent message
Task difficulty	Item response theory (IRT)
Group/individual differences	Multilevel analysis (ML) (hierarchical linear models, HLM)
Pivotal event	Breakpoint analysis
Time periods	Breakpoint analysis and multilevel analysis
Multiple target events Model latent processes underlying events, indirect, mediation effects and measurement error	Multilevel structural equation modeling (ML-SEM)
Later group/individual outcomes	Add outcome and its interaction as explanatory variables and multilevel moderation via random effects

(a) stored parent message, (b) item response theory, (c) multilevel analysis, (d) breakpoint analysis, (e, f, g, h) multilevel structural equation model, and (i) multilevel moderation via random effects.

3.1.1 Parallel Chats and Trees

Although much talk occurs in sequence with one speaker after another, sometimes learners separate into *parallel conversations* or online participants engage with messages according to their thread structure (often *trees*) rather than in temporal order (Chiu and Lehmann-Willenbrock 2016). To analyze such nonsequential data, researchers can identify and store the previous message of each message in a variable *parent message*; specifically, a computer program can create this variable by traversing parallel chats/conversations or trees of messages/turns of talk (this program is available in Chen and Chiu 2008).

3.1.2 Task Difficulty

Tasks differ in difficulty, so ignoring these differences can mask a student's learning progress (or difficulties). *Item response theory* simultaneously models the difficulty of each task and each student's overall competence (along with guessing success on multiple choice questions, Embretson and Reise 2013). An IRT model that incorporates a time parameter enables modeling of learning (or changes across time, *additive factors model*, Cen et al. 2006).

3.1.3 Group/Individual Differences

Groups and individuals likely differ. Specifically, messages written by the same student likely resemble one another more than those by different students. Likewise, messages in the same thread/topic likely resemble one another more than those in different threads/topics. CP and SA cannot model these differences, and a regression would negatively bias the standard errors. Hence, we apply a *multilevel analysis* to yield unbiased results (Goldstein 2011; also known as *hierarchical linear modeling*, Bryk and Raudenbush 1992). In general, such nested data (students within groups within classrooms within schools, etc.) require multilevel analysis for accurate results (Goldstein 2011).

3.1.4 Differences Across Time

An outcome (e.g., explanation) might be more likely at the beginning, the middle, the end, or in a specific time interval (Chiu and Lehmann-Willenbrock 2016). Furthermore, the relations among explanatory variables and outcomes might differ

across time (Chiu and Lehmann-Willenbrock 2016). Although humans can decide how to divide a stream of data into time periods, past studies show that such subjective methods are unreliable (e.g., Wolery et al. 2010).

In contrast, *breakpoint analysis* objectively identifies *pivotal events* that substantially increase (or decrease) the likelihood of an outcome (e.g., explanation, Chiu and Lehmann-Willenbrock 2016). Researchers can then test explanatory models to characterize when these pivotal events occur. For example, *discussion summaries* were often breakpoints that sharply elevated the quality of online discussions, and students assigned the roles of *synthesizer* or *wrapper* were far more likely than others to create discussion summaries (Wise and Chiu 2011).

These pivotal events divide the data series into distinct *time periods* of significantly higher versus lower likelihoods of the outcome (e.g., explanations are much more likely in one time period than another, Chiu and Lehmann-Willenbrock 2016). These time periods provide an additional level to the above multilevel analysis (Chiu and Lehmann-Willenbrock 2016). Researchers can then test whether relations among variables are stronger in some time periods than in others. For example, when groups of high school students worked on an algebra problem, a correct evaluation of a groupmate's idea raised the likelihood of a correct contributions in most time periods, but not all of them; the effect ranged from −0.3% to +9% across time periods (Chiu 2008).

3.1.5 Multiple Target Events, Latent Process, Indirect Effect, and Measurement Error

Often, researchers are interested in how processes affect multiple types of targeted events (e.g., explanations and correct, new ideas [*micro-creativity*], Chiu and Lehmann-Willenbrock 2016). As multiple types of target events might be related to one another, standard analyses designed for a single dependent variable can yield biased standard errors (Kennedy 2008). Hence, researchers have developed methods such as *multilevel structural equation models* (ML-SEM, Joreskog and Sorbom 2015) that simultaneously test multiple dependent variables; in the above algebra group, problem-solving example, a justification might yield both another justification and micro-creativity (*justification [−1] → justification*; *justification [−1] → micro-creativity*). ML-SEMs also properly test indirect mediation effects [$X \rightarrow M \rightarrow Y$] and combine multiple measures of a single construct into a single index that increases precision, such as tests to measure intelligence (Muthén and Muthén 2018); continuing with the algebra group problem-solving example, for example, a correct evaluation often followed by another correct evaluation, which in turn is followed by micro-creativity (*correct evaluation [−2] → correct evaluation [−1] → micro-creativity*.

3.1.6 Later Group/Individual Outcomes

In addition to the immediate consequences of processes on target events, researchers are often interested in whether such sequences have longer term effects, such as the quality of a group's final solution to the current problem or later individual test scores (Chiu 2018). The traditional approach of aggregating event-level data to the individual or group level (or any higher level) discards substantial information and yields inaccurate results (Goldstein 2011).

Instead, researchers can use an event-level analysis to utilize all the available data (Chiu 2018). Consider groups of students designing plans to reduce climate change (e.g., reduce cafeteria beef dishes to reduce cow methane). A researcher wants to know if a group that has more *disagree* → *explain* sequences than others creates a superior *group plan*. Chiu (2018) showed how to test this hypothesis via a regression with the dependent variable *explain* and the following explanatory variables: *disagree [−1]*, *group plan*, and the interaction term *disagree [−1]* × *group_plan*. This message-level specification asks, "In groups with higher plan scores, is a *disagree* message more likely to be followed by an *explain* message?" The message sequences occur before the group plan, and time cannot flow backward, so the group plan cannot influence the message sequences.

Likewise, a researcher can also test whether individuals that participate in more *disagree* → *explain* sequences than others have higher subsequent science test scores by adding the following explanatory variables to the above regression specification: *test score* and *disagree [−1]* × *test score*. Hence, this elaborated specification simultaneously tests whether *disagree* → *explain* sequences link to group plans or individual science test scores. More generally, a regression does not mathematically dictate the direction of causality, so traditional outcomes can serve as independent variables (Chiu 2018). For nested data (e.g., messages within time periods, see above), modeling such interactions requires a *multilevel moderation* via *random effects* (Chiu and Lehmann-Willenbrock 2016).

In short, statistical methods enable researchers to test hypotheses regarding sequences of events, their antecedents at any level, parallel chats and trees, task difficulty differences, group/individual differences, pivotal events, time periods, multiple target events, latent processes, indirect links, measurement error, and later group/individual outcomes. See Chiu and Lehmann-Willenbrock's (2016) *statistical discourse analysis* (SDA) regarding integration of most of the above analyses, along with statistical methods for addressing related issues (e.g., missing data, inter-rater reliability, false positives, etc.).

3.2 Recent Developments in Stochastic Modeling

3.2.1 Extensions of Hidden Markov Models

In CSCL research, HMMs have been mainly applied for practical purposes: to provide a compact representation of long interaction sequences, one that is useful for making predictions. Learning is reflected not only in talk and conversation, but also in eye gaze, movement, and gestures. HMMs can be used on such kinds of data as well. This has been made easier as recent years have seen a vast expansion of the use of HMMs, enabled by the introduction of software packages that remove constraints on data modeling. Focusing on what is available in R, HMM packages have been developed that can learn from multiple channel observation sequences (Visser and Speekenbrink 2010), relevant for instance for cases where eye-tracking is combined with observations of interaction and verbal data (Schneider et al. 2018). In the same package, covariates can be added for initial and transition probabilities. This allows us, for instance, to model cases in which the participants are provided with time-dependent additional information, such as observations on a peer tutor (Walker et al. 2014). Another important extension concerns multiple observation sequences: the hidden states are seen as representing a distribution of states (O'Connell and Højsgaard 2011; Turner and Liu 2014). There are also extensions for modeling continuous time processes (Jackson 2011), relevant in CSCL for research that includes, for instance, physiological measurements (Mandryk and Inkpen 2004). One of the most comprehensive HMM packages for R currently available that reflects a range of these extensions is seqHMM (Helske and Helske 2017).

3.2.2 Dynamic Bayesian networks

An important development in stochastic modeling of temporal processes is dynamic Bayesian networks (DBNs). They provide a perspective for probabilistic reasoning over time that unifies (hidden) Markov modeling in all its variants with the Bayesian approach to modeling diagnostic reasoning, decision-making, and measurement.

The ontology of a DBN is such that the world is a series of snapshots—of time slices—each of which contains a number of (unobservable) state variables and a number of observable variables that are indicators for states. For the simplest case of a DBN, we assume that the variables and their links are exactly replicated from slice to slice and that the DBN itself represents a first-order Markov process: each variable has "parents" (is linked to) only in its own slice and/or the immediately preceding slice (Russell and Norvig 2016, p. 590).

To provide an example, the study on math learning by peer tutoring described in (Bergner et al. 2017) uses an input–output HMM to model the relation between tutor input, tutee's capability (the hidden state), and the correctness of observed tutee actions (see Fig. 2). This model makes it explicit that a capability increase on side of

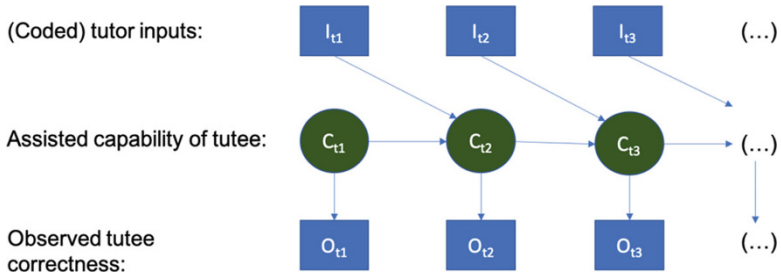

Fig. 2 DBN model of learning in a tutorial dialogue after Bergner et al. (2017)

the tutee depends on the tutor as well as the tutee. Linking the variables of a system makes it computationally much more tractable than when only the set of variables is provided. Because a DBN can express more structure than an HMM, it becomes more efficiently computable and it allows expression of a wider set of theoretical assumptions.

As regards software support, the bnlearn package in R, for instance, can be used to construct DBNs theory-driven or to learn them from data (Nagarajan et al. 2013).

4 The Future

On the horizon are dynamic social network analysis, massive data, automatic analyses, and qualitative/quantitative analysis cycles. While *social network analyses* can examine attributes of fixed networks of learners or ideas (*epistemic network analysis*), *dynamic social network analysis* offers the promise of examining how networks of learners, ideas, or both change over time (Oshima et al. 2018; Sarkar and Moore 2006; Shaffer et al. 2009).

Massive data (colloquially, big data) encompass sharply greater volume, complexity, and velocity (e.g., from massive, open, online courses or MOOCs; National Research Council 2013). The increasing addition of computer chips into objects around us (colloquially, internet of things) and their technological embrace by educators to aid student learning is creating voluminous amounts of electronic data (Picciano 2012). Greater volumes of data largely enhance statistical analyses and enable greater precision in the results (Cohen et al. 2003). Although some data are in the familiar form of numbers, much of it is text, images, or videos (Gandomi and Haider 2015). These data require substantial effort before conversion into numbers for statistical analyses (e.g., 1 for presence of an image vs. 0 for its absence), so collaborations among experts in computational linguistics, image processing, acoustics, and statistics will likely become necessary (Bello-Orgaz et al. 2016). Also, high-velocity data collection entails repeated dynamic analyses to yield updated results (each day, hour, minute, etc.; Zikopoulos and Eaton 2011).

The growing size, complexity, and velocity of massive data and the accompanying demand for comprehensive, nuanced, updated analyses of them exceed human capacity, so they motivate automated, computer programs to take over increasingly greater statistics responsibilities (Assunção et al. 2015). After computer programs informed by computational linguistics, image processing, and acoustics create the databases (Bello-Orgaz et al. 2016), *artificial intelligence expert systems* can select and run the statistical analyses (repeatedly for high-velocity incoming data), interpret the results, and produce reports for humans (Korb and Nicholson 2010).

The automation of statistical analyses also frees up human time for detailed qualitative analyses, so that both analyses mutually inform each other's subsequent analyses, provide mutually supportive evidence, and complement each other's strengths and weaknesses (Teddlie and Tashakkori 2009). For example, an initial, qualitative case study can select and scrutinize important phenomena in context to develop theory by identifying constructs, operationalizing them, recognizing patterns, and specifying hypotheses (possibly aided by data mining, Feldman and Sanger 2007). Next, a statistical analysis tests these hypotheses, identifies pivotal breakpoints, and pinpoints instances in the data that fit the theory extremely well or extremely poorly (Chiu 2013). The hypothesis testing results, breakpoints, well-fit instances, and poorly fitting instances target specific data for another round of qualitative analysis (Chiu 2013). This qualitative analysis can refine the hypotheses, develop new ones for breakpoints or poorly fitting instances for another round of statistical analyses, and so on (Teddlie and Tashakkori 2009). Researchers can flexibly start or stop at any point in the above multistep qualitative/quantitative cycle (Chiu 2013).

Also interesting for bringing qualitative and quantitative approaches into closer contact are *deterministic process models*. Deterministic modeling applies when the structure of the process is known and one is interested in the behavior of the process under certain conditions. Deterministic models are particularly relevant for CSCL research when processes are *designed*, such as for the study of collaboration and argumentation scripts (Weinberger and Fischer 2006).

A deterministic process differs from a fixed, invariant sequence of steps (activities, events). With known start and end states, it is a finite set of both states and activities (or actions) that can yield to an *infinite* number of different event sequences (e.g., chess). Computer science and operations research have examined deterministic process models expressed in forms such as finite-state machines and Petri Nets (Reimann 2009). As these models can represent choice and parallelism, they can help answer questions such as: Is an observed, sequence of events alignable with a particular designed process? Given a set of sequences of events, can a single deterministic model describe them?

Deterministic models are also relevant in situations where the learners involved have knowledge about the process as a whole; for instance, participants in a formal discussion know the "moves" allowed as well as the end state (Schwarz and Baker 2016). We can therefore assume that their behavior in the discussion will to some extent be guided by this knowledge, by a sense of well-formedness. In human affairs, such situations abound, from social conduct in general to work processes. Although

obviously relevant for CSCL research, applications have been rare so far (Bannert et al. 2014; Reimann et al. 2009). The same can be said about the type of deterministic models that represent knowledge and beliefs of individual agents and simulate the interaction with other agents and resources, such as agent-based models. While of high relevance to phenomena studied in CSCL, applications are very rare. To appreciate the role they could play, Abrahamson et al.'s model of stratification of learning zones in the collaborative (math) classroom provides an excellent example (Abrahamson et al. 2007).

In conclusion, a wide range of methods for analyzing and modeling temporal data is available to CSCL researchers, ranging from stochastic and statistical to deterministic computational. Our recommendation is to embrace the notions of model and modeling to a much deeper and much more comprehensive extent than has been the case in the past, by exploiting the potential that lies in combining theoretical with empirical modeling. We hope this chapter will make a small contribution to this widening of minds.

References

Abrahamson, D., Blikstein, P., & Wilensky, U. (2007). Classroom model, model classroom: Computer-supported methodology for investigating collaborative-learning pedagogy. In C. Chinn, G. Erkens, & S. Puntambekar (Eds.), Proceedings of the 8th international conference on computer supported collaborative learning (CSCL) (Vol. 8, part 1, pp. 49–58). International Society of the Learning Sciences.

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A., & Buyya, R. (2015). Big data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing, 79*, 3–15.

Bakeman, R., & Gottman, J. M. (1986). *Observing interaction: An introduction to sequential analysis*. Cambridge: Cambridge University Press.

Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition and Learning, 9*(2), 161–185.

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45–59.

Bergner, Y., Walker, E., & Ogan, A. (2017). Dynamic Bayesian network models for peer tutoring interactions. In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 249–268). New York: Springer.

Boyer, K. E., Ha, E. Y., Phillips, R., Wallis, M. D., Vouk, M. A., & Lester, J. (2009). Inferring tutorial dialogue structure with hidden Markov modeling. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications—EdAppsNLP '09* (pp. 19–26). Association for Computational Linguistics.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. London: Sage.

Cen, H., Koedinger, K., & Junker, B. (2006). Learning factors analysis–a general method for cognitive model evaluation and improvement. In M. Ikeda, K. D. Ashley, & T. W. Chan (Eds.), *Intelligent tutoring systems, lecture notes in computer science* (Vol. 4053, pp. 164–175). New York: Springer.

Chen, G., & Chiu, M. M. (2008). Online discussion processes: Effects of earlier messages' evaluations, knowledge content, social cues and personal information on later messages. *Computers and Education, 50*, 678–692.

Chiu, M. M. (2008). Flowing toward correct contributions during groups' mathematics problem solving: A statistical discourse analysis. *Journal of the Learning Sciences, 17*(3), 415–463. https://doi.org/10.1080/10508400802224830.

Chiu, M. M. (2013). Cycles of discourse analysis <=> statistical discourse analysis. In *10th International conference on computer supported collaborative learning*, Madison, WI, USA.

Chiu, M. M. (2018). Statistically modelling effects of dynamic processes on outcomes: An example of discourse sequences and group solutions. *Journal of Learning Analytics, 5*(1), 75–91.

Chiu, M. M., & Lehmann-Willenbrock, N. (2016). Statistical discourse analysis: Modeling sequences of individual behaviors during group interactions across time. Group Dynamics: Theory, Research, and Practice, 20(3), 242–258. DOI: 10.1037/gdn0000048

Cohen, J., West, S. G., Aiken, L., & Cohen, P. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum.

Cress, U. (2008). The need for considering multilevel analysis in CSCL research—an appeal for the use of more advanced statistical methods. *International Journal of Computer-Supported Collaborative Learning, 3*, 69–84.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Hove, East Sussex, UK: Psychology Press.

Farran, D. C., & Son-Yarbrough, W. (2001). Title I funded preschools as a developmental context for children's play and verbal behaviors. *Early Childhood Research Quarterly, 16*(2), 245–262.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge: Cambridge University Press.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.

Goldstein, H. (2011). *Multilevel statistical models*. London: Edward Arnold.

Gottman, J. M., & Roy, A. K. (1990). *Sequential analysis: A guide for behavioral researchers*. Cambridge: Cambridge University Press.

Greene, W. H. (1997). *Econometric analysis* (3rd ed.). London: Prentice-Hall.

Helske, S., & Helske, J. (2017). *Mixture hidden Markov models for sequence data: The seqHMM package in R*. Retrieved from http://arxiv.org/abs/1704.00543

Jackson, C. H. (2011). Multi-state models for panel data: The msm package for R. *Journal of Statistical Software, 38*(8), 1–29.

Joreskog, K., & Sorbom, D. (2015). *LISREL 9.2*. New York: Scientific Software International.

Kennedy, P. (2008). *Guide to econometrics*. New York: Wiley-Blackwell.

Korb, K. B., & Nicholson, A. E. (2010). *Bayesian artificial intelligence*. Boca Raton, FL: CRC Press.

Loehlin, C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis*. Hove, East Sussex, UK: Psychology Press.

Mandryk, R. L., & Inkpen, K. M. (2004). Physiological indicators for the evaluation of co-located collaborative play. In *Proceedings of the 2004 ACM conference on Computer Supported Cooperative Work—CSCW '04* (pp. 102–111). Association for Computing Machinery.

Muthén, L. K., & Muthén, B. O. (2018). *Mplus 8.1*. Los Angeles, CA: Muthén & Muthén.

Nagarajan, R., Scutari, M., & Lèbre, S. (2013). *Bayesian networks in R*. New York: Springer.

National Research Council. (2013). *Frontiers in massive data analysis*. Washington, DC: National Academies Press.

O'Connell, J., & Højsgaard, S. (2011). Hidden semi Markov models for multiple observation sequences: The mhsmm package for R. *Journal of Statistical Software, 39*(4), 1–22.

Oshima, J., Oshima, R., & Fujita, W. (2018). A mixed-methods approach to analyze shared epistemic agency in jigsaw instruction at multiple scales of temporality. *Journal of Learning Analytics, 5*(1), 10–24.

Picciano, A. G. (2012). The evolution of big data and learning analytics in American higher education. *Journal of Asynchronous Learning Networks, 16*(3), 9–20.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

Reimann, P. (2009). Time is precious: Variable- and event-centred approaches to process analysis in CSCL research. *International Journal of Computer-Supported Collaborative Learning, 4*, 239–257.
Reimann, P., Frerejean, J., & Thompson, K. (2009). Using process mining to identify models of group decision making processes in chat data. In C. O'Malley, D. Suthers, P. Reimann, & A. Dimitracopoulou (Eds.), *Computer-supported collaborative learning practices: CSCL2009 conference proceedings* (pp. 98–107). International Society for the Learning Sciences.
Russell, S., & Norvig, P. (2016). *Artificial intelligence: A modern approach* (global edition). London: Prentice-Hall.
Sarkar, P., & Moore, A. W. (2006). Dynamic social network analysis using latent space models. In Y. Weiss, B. Scholkopf, and J. Platt (Eds.) *Advances in neural information processing systems 18* (pp. 1145–1152). Cambridge, MA: MIT Press.
Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., & Pea, R. (2018). Leveraging mobile eye-trackers to capture joint visual attention in co-located collaborative learning groups. *International Journal of Computer-Supported Collaborative Learning, 13*(3), 241–261.
Schwarz, B., & Baker, M. (2016). *Dialogue, Argumentation and education*. Cambridge: Cambridge University Press.
Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., Frank, K., Rupp, A. A., & Mislevy, R. (2009). Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media, 1*(2), 33–53.
Soller, A. (2004). Computational modeling and analysis of knowledge sharing in collaborative distance learning. *User Modeling and User-Adapted Interaction, 14*, 351–381.
Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. London: Sage.
Tuckman, B. W. (1965). Developmental sequence in small groups. *Psychological Bulletin, 63*(6), 384–399.
Turner, R., & Liu, L. (2014). *Hmm.discnp: Hidden Markov models with discrete non-parametric observation distributions. R Package Version 0.2-3*. Retrieved from http://CRAN.R-project.org/package=hmm.discnp
Visser, I., & Speekenbrink, M. (2010). depmixS4: An R Package for Hidden Markov Models. *Journal of Statistical Software, 36*, 1–21.
Walker, E., Rummel, N., & Koedinger, K. R. (2014). Adaptive intelligent support to improve peer tutoring in algebra. *International Journal of Artificial Intelligence in Education, 24*(1), 33–61.
Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education, 46*(1), 71–95.
Wise, A., & Chiu, M. M. (2011). Analyzing temporal patterns of knowledge construction in a role-based online discussion. *International Journal of Computer-Supported Collaborative Learning, 6*, 445–470.
Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education, 44*(1), 18–28.
Zikopoulos, P., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class Hadoop and streaming data*. New York: McGraw-Hill Osborne Media.

Further Readings

Abrahamson, D., Blikstein, P., & Wilensky, U. (2007). Classroom model, model classroom: Computer-supported methodology for investigating collaborative-learning pedagogy. In C. Chinn, G. Erkens, & S. Puntambekar (Eds.), *Proceedings of the eighth International Conference on Computer Supported Collaborative Learning* (CSCL) (Vol. 8, Part 1, pp. 49–58). International Society of the Learning Sciences. A powerful demonstration of how

(deterministic) computational modeling can interact with empirical (classroom) research. Using the agent-based modeling tool, NetLogo, the authors provide an analysis of the mechanisms that lead to the emergence of stratified learning zones in a prototypical collaborative classroom activity. Also important because it highlights the tension between collaborative solving problems and learning from collaboration.

Bergner, Y., Walker, E., & Ogan, A. (2017). Dynamic Bayesian Network models for peer tutoring interactions. In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 249–268). Springer. This chapter provides a nice illustration of the use of modern HMM approaches to analyzing (peer) tutorial dialogue. While an important area of collaborative learning, research on tutor–tutee dialogue is only partially reflected in the CSCL literature, with this chapter providing a welcome connection between CSCL, AI in Education, and assessment research. It includes an application in the context of an empirical study.

Chiu, M. M. (2008). Flowing toward correct contributions during groups' mathematics problem solving: A statistical discourse analysis. *Journal of the Learning Sciences, 17*(3), 415–463. This empirical study applied statistical discourse analysis to test whether (a) groups that created more correct, new ideas (micro-creativity) were more likely to solve a problem and (b) students' recent actions (microtime context of evaluations, questions, justifications, politeness, and status differences) increased subsequent micro-creativity.

Chiu, M. M., & Lehmann-Willenbrock, N. (2016). Statistical discourse analysis: Modeling sequences of individual behaviors during group interactions across time. *Group Dynamics: Theory, Research, and Practice, 20*(3), 242–258. This article showcases statistical discourse analysis, a method that integrates most of the above methods (parallel chats, trees, group/individual differences, pivotal events, time periods, multiple target events, indirect effects, later group outcomes) and addresses related issues (e.g., missing data, inter-rater reliability, false positives, etc.).

Reimann, P. (2009). Time is precious: Variable- and event-centred approaches to process analysis in CSCL research. *International Journal of Computer-Supported Collaborative Learning, 4*, 239–257. This methodological paper provides an overview of qualitative, quantitative, and computational methods for analyzing temporal data in CSCL. It argues that there is a rather fundamental difference between explaining collaboration over time in terms of variables versus explaining them in terms of events. Implications for doing temporal analysis are discussed.

NAPLES Video

Chiu, M. M. (2018). How to statistically model processes? Statistical discourse analysis. *Network of Academic Programs in the Learning Sciences* (*NAPLeS*) *webinar*. http://isls-naples.psy.lmu.de/intro/all-webinars/chiu/index.html