the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2020), 73, 474–505

www.wileyonlinelibrary.com

A latent topic model with Markov transition for process data

Haochen Xu[1]*, Guanhua Fang[2] and Zhiliang Ying[2]
[1]Fudan University, Shanghai, China
[2]Columbia University, New York, New York, USA

We propose a latent topic model with a Markov transition for process data, which consists of time-stamped events recorded in a log file. Such data are becoming more widely available in computer-based educational assessment with complex problem-solving items. The proposed model can be viewed as an extension of the hierarchical Bayesian topic model with a hidden Markov structure to accommodate the underlying evolution of an examinee's latent state. Using topic transition probabilities along with response times enables us to capture examinees' learning trajectories, making clustering/classification more efficient. A forward-backward variational expectation-maximization (FB-VEM) algorithm is developed to tackle the challenging computational problem. Useful theoretical properties are established under certain asymptotic regimes. The proposed method is applied to a complex problem-solving item in the 2012 version of the Programme for International Student Assessment (PISA).

1. Introduction

Testing examinees' complex problem-solving (CPS) ability is becoming of primary interest in computer-based assessments. Recently, a number of prominent large-scale educational assessments have included CPS items as their essential components; see, for example, the 2012, 2015 and 2018 versions of the Programme for International Student Assessment (PISA; OECD, 2014, 2016), 2012 Programme for International Assessment of Adult Competencies (PIAAC; Goodman, Finnegan, Mohadjer, Krenzke, & Hogan, 2013), and Assessment and Teaching of 21st Century Skills (ATC21S; Griffin, McGaw, & Care, 2012). The CPS items in these tests are interaction-oriented, requiring students to react to new information adaptively as they receive it. A CPS item typically asks examinees to solve a problem in a simulation environment. In order to arrive at correct answers, examinees may need to learn the environment and acquire knowledge sequentially and interactively.

For an examinee, the process of solving a CPS item (i.e., the examinee's action sequence) is recorded as a log file. These log file data are commonly known as *process data*. Although traditional psychometric models and statistical methods are not directly applicable, there is a growing literature on process data with varying foci. Fischer, Greiff, and Funke (2011) reviewed the history of CPS in a variety of research domains and emphasized the importance of information reduction, model building and evaluation in CPS data analysis. Halpin and De Boeck (2013) proposed the use of the Hawkes process to

*Correspondence should be addressed to Haochen Xu, Fudan University, Room 1904, East Guanghua Main Building, 220 Handan Road, Shanghai 200433, China (email: hcxu15@fudan.edu.cn).

DOI:10.1111/bmsp.12197

model interactions among examinees in collaborative CPS items. He and von Davier (2015, 2016) pursued a similar goal by grouping consecutive events into *n*-grams and measuring their association with the outcomes. He, von Davier, Greiff, Steinhauer, and Borysewicz (2017) discussed the issues and challenges associated with measurement of collaborative problem-solving skills by using an example in PISA 2015. Polyak, von Davier, and Peterschmidt (2017) presented an application of computational psychometrics to collaborative CPS items in the form of continuous Bayesian evidence tracing. The behavioural paths in an online collaborative problem-solving item are studied by Vista, Care, and Awwal (2017) through event transition graphs. Xu, Fang, Chen, Liu, and Ying (2018) used the latent class model to cluster population based on event histories and response times. Qiao and Jiao (2018) adopted various classification methods for a data set from PISA 2012 to achieve a better accuracy. Chen, Li, Liu, and Ying (2019) focused on predicting success probability and average residual time for task completion.

Despite these efforts, the modelling and analysis of process data are still in their infancy. Most approaches are *ad hoc* in nature and there is a lack of consensus as to how to develop a comprehensive approach which can handle a large variety of process data. It is desirable to provide a statistical framework that can summarize and handle the important features of process data, that is, event types and timing (sequence), individual and event type heterogeneity, and have parameters with meaningful psychometric interpretation.

In this paper, we propose a hierarchical statistical model with a Markov structure to characterize both the order/type of events and individual-level effects. Within this framework, we model the event sequence or process data through a latent Markov chain which represents the evolving latent profiles of the examinee. We assume the first event of the test taker follows some common (baseline) initial distribution. Later events then evolve by following a Markov chain with person-specific transition probabilities and person-specific gap time distributions. We assign a latent topic to each event, with number of topics much smaller than the number of event types, allowing it to have a potential meaning. For computation, a known challenging issue in such modelling, we propose a new method, combining the forward-backward algorithm, variational Bayesian method and expectation-maximization (EM) algorithm. Both theoretical and simulation results show that the new method not only is computational tractable, but also provides reasonably good parameter estimation.

The remainder of this paper is organized as follows. In Section 2 we introduce notation and give the model specification. In Section 3 we present a new forward-backward variational expectation-maximization (FB-VEM) algorithm to obtain parameter estimation. In Section 4 we establish some theoretical properties for the proposed estimators. In Section 5 we summarize our simulation results. In Section 6 we apply the proposed method to the process data from a CPS item in PISA 2012. Section 7 concludes.

2. Latent topic analysis with Markov transition

2.1. Notation and setting

Recall that the log file of an examinee contains a sequence of ordered events (actions) coupled with time-stamps. We use N to denote the total number of events over testing period $[0, \tau]$, where τ is the termination time. The observed data sequence for this examinee is denoted by $\{(e_1, t_1), \ldots, (e_n, t_n), \ldots, (e_N, t_N)\}$, where e_n is the nth event and t_n is the corresponding time-stamp. Here e_n takes a value from set ε which consists of all distinct event types. We use V to denote the cardinality of ε. For notational simplicity, we let $\mathbf{t}_{1:N} = \{t_n: n = 1, \ldots, N\}$ be the set of ordered event times, where $t_0 = 0 < t_1 < \ldots < t_N = \tau$. Also let $\mathbf{e}_{1:N} = \{e_n: n = 1, \ldots, N\}$ be the set of corresponding events. Below, we use the log file of the climate control question in PISA 2012 as an example to illustrate the process data structure.

The climate control question is a problem-solving item from PISA 2012. Around 510,000 15-year-old students from over 60 countries and economies completed the PISA assessment in 2012. Among them, approximately 85,000 students took the problem-solving tests. As seen in Figure 1, the climate control item gives examinees a new air conditioner and asks them to connect three controls to temperature and/or humidity. They can explore the top, central and bottom controls by moving the corresponding sliders and clicking the ‘apply’ or ‘reset’ button. After they have clicked, the temperature and humidity levels are updated in the panel. Once they finish exploring, the examinees need to answer the question, that is, to draw lines in the diagram (Figure 2), connecting the sliders to temperature/humidity.

Table 1 contains the log data of one examinee’s event history. This examinee performs 12 actions in 88 s. The ‘time’ column gives the recorded times (in seconds) at which the 12 actions were taken. The top, central and bottom setting columns indicate the positions of the three sliders. Detailed explanations of all columns are given in Table 5 in the Appendix S1. Because the positions of the sliders are updated only when the ‘apply’ button is clicked, we will only consider those events and event times for which the corresponding event type is ‘apply’. As a result, the set of all distinct events becomes $\varepsilon = \{(0, 0, 0), \ldots, (2, 2, 2)\}$, which is the set of all

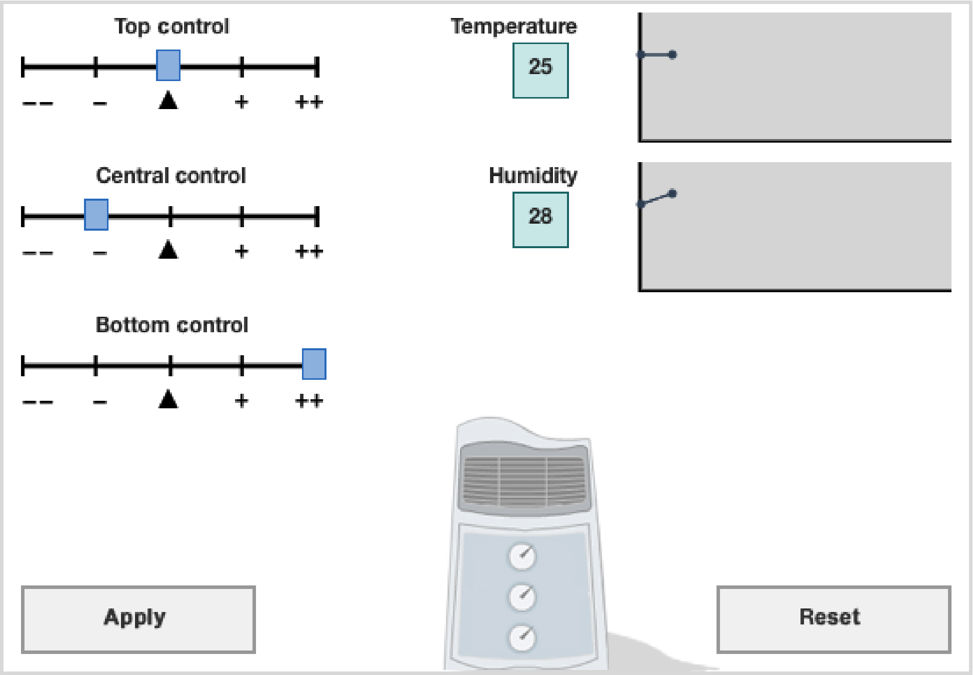

Figure 1. The climate control item in Programme for International Student Assessment (PISA) 2012. [Colour figure can be viewed at wileyonlinelibrary.com]

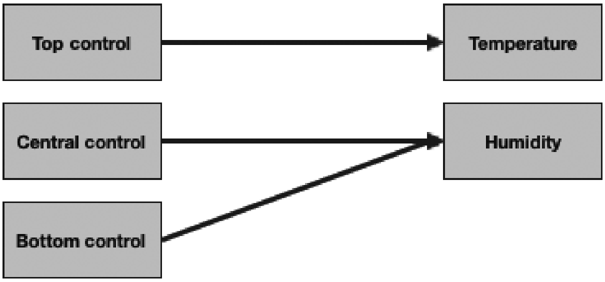

Figure 2. The climate control item answer diagram.

combinations of three slider positions. After removing unused rows (1–3, 5 ,7–8, 10–12), the cleaned data sequence for this examinee becomes {e_1 = (2, 0, 0), e_2 = (0, 2, 0), e_3 = (0, 2, 2); t_1 = 60.1, t_2 = 70.0, t_3 = 80.5}. Details of the data cleaning can be found in Section 6.

2.2. Model specification

To define our model, we first introduce a (latent) topic sequence, denoted by $\mathbf{z}_{1:N} = \{z_1, \ldots, z_n, \ldots, z_N\}$. We assume $z_n \in \mathcal{Z} = \{1, \ldots, K\}$, with K as the number of latent topics. In general, we can write the density function of the observed data as

$$p(\mathbf{e}_{1:N}, \mathbf{t}_{1:N}) = \sum_{\mathbf{z}_{1:N}} \left[\prod_{n=1}^{N} p(e_n, t_n | \mathbf{e}_{1:(n-1)}, \mathbf{t}_{1:(n-1)}, z_{1:N}) \right] p(\mathbf{z}_{1:N}), \tag{1}$$

where, for notational simplicity, we let e_0 and t_0 to denote empty event and time respectively, such that$p(e_1, t_1|e_0, t_0, \mathbf{z}_{1:N}) = p(e_1, t_1|\mathbf{z}_{1:N})$. Assuming that the nth event (e_n, t_n) depends only on the topic transitions from z_{n-1} to z_n and its preceding time-stamp t_{n-1}, we have

$$p(e_n, t_n | \mathbf{e}_{1:(n-1)}, \mathbf{t}_{1:(n-1)}, \mathbf{z}_{1:N}) = p(e_n, t_n | t_{n-1}, z_{n-1}, z_n). \tag{2}$$

We further assume that e_n and t_n are conditionally independent given $z_{n\text{-}1}$ and z_n, and the right-hand side of (2) becomes

$$p(e_n, t_n | t_{n-1}, z_{n-1}, z_n) = p(e_n | z_n) p(t_n | t_{n-1}, z_{n-1}, z_n). \tag{3}$$

Finally, we assume the latent topic sequence $\{z_n\}_{n=1}^{N}$ is a Markov chain, that is, $p(\mathbf{z}_{1:N}) = \prod_{n=1}^{N} p(z_n | z_{n-1})$, where $p(z_1 | z_0) = p(z_1)$. Under these assumptions, (1) becomes

$$p(\mathbf{e}_{1:N}, \mathbf{t}_{1:N}) = \sum_{\mathbf{z}_{1:N}} \left[\prod_{n=1}^{N} p(e_n | z_n) p(t_n | t_{n-1}, z_{n-1}, z_n) \right] \prod_{n=1}^{N} p(z_n | z_{n-1}). \tag{4}$$

We specify the probability distributions on the right-hand side of (4) as follows:

Table 1. Log data of an examinee's process of solving the climate control item

Event number	Event	Time	Event type	Top setting	Central setting	Bottom setting	Temp. value	Humid. value	Diag. state
1	START_ITEM	0.00	NULL	NULL	NULL	NULL	NULL	NULL	NULL
2	ACER_EVENT	40.60	Diagram	NULL	NULL	NULL	NULL	NULL	000000
3	ACER_EVENT	42.60	Diagram	NULL	NULL	NULL	NULL	NULL	000000
4	ACER_EVENT	60.10	Apply	2	0	0	29	25	NULL
5	ACER_EVENT	65.00	Diagram	NULL	NULL	NULL	NULL	NULL	100000
6	ACER_EVENT	70.00	Apply	0	2	0	29	27	NULL
7	ACER_EVENT	76.40	Diagram	NULL	NULL	NULL	NULL	NULL	100000
8	ACER_EVENT	77.20	Diagram	NULL	NULL	NULL	NULL	NULL	100100
9	ACER_EVENT	80.50	Apply	0	2	2	29	33	NULL
10	ACER_EVENT	84.60	Diagram	NULL	NULL	NULL	NULL	NULL	100100
11	ACER_EVENT	85.10	Diagram	NULL	NULL	NULL	NULL	NULL	100101
12	END_ITEM	88.00	NULL	NULL	NULL	NULL	NULL	NULL	NULL

$$e_n | z_n = k \sim \text{Multinomial}(\mathbf{b}_k), \quad \mathbf{b}_k = \left(b_{k,1}, \ldots, b_{k,V}\right), \tag{5}$$

$$z_n | z_{n-1} = k' \sim \text{Multinomial}\left(\boldsymbol{\lambda}^{k'}\right), \quad \boldsymbol{\lambda}^{k'} = \left(\lambda_1^{k'}, \ldots, \lambda_K^{k'}\right), \tag{6}$$

$$z_1 \sim \text{Multinomial}\left(\mathbf{p}^0\right), \tag{7}$$

$$t_n - t_{n-1} | z_{n-1} = k', z_n = k, \xi, G \sim \text{Exponential}(\xi e^{g_{k',k}}), \tag{8}$$

$$\xi | a, d \sim \text{Gamma}(a, d). \tag{9}$$

Furthermore, the random matrix $\Lambda \equiv \left(\boldsymbol{\lambda}^1, \ldots, \boldsymbol{\lambda}^K\right)$ is assumed to follow a Dirichlet prior with parameter $R = \left(\mathbf{r}^1, \ldots, \mathbf{r}^K\right)^T$ such that

$$\boldsymbol{\lambda}^{k'} \sim \text{Dir}\left(\mathbf{r}^{k'}\right), \quad \mathbf{r}^{k'} = \left(r_1^{k'}, \ldots, r_K^{k'}\right). \tag{10}$$

In view of (4)–(10), we have

$$\begin{aligned} p(\mathbf{e}_{1:N}, \mathbf{t}_{1:N}) = \int\limits_{\Lambda, \xi} & \left\{ \sum_{\mathbf{z}_{1:N}} \prod_{n=1}^{N} p(e_n | z_n, B) p(t_n | t_{n-1}, z_{n-1}, z_n, \xi, G) p(z_n | z_{n-1}, \Lambda) \right\} \\ & \cdot p(\xi | a, d) p(\Lambda | R) \mathrm{d}\xi \mathrm{d}\Lambda, \end{aligned} \tag{11}$$

where $B \equiv (\mathbf{b}_1, \ldots, \mathbf{b}_K)$ and $G \equiv (g_{k',k})_{K \times K}$ are model parameters.

By definition, B is a $K \times V$ matrix that connects the observed event types to latent topics, and thus may be interpreted as ‘factor loadings’. It is at population level that does not vary among different examinees. On the other hand, Λ varies with different examinees. Thus, for a particular examinee, the corresponding Λ may be viewed as a personal transition probability matrix. The intensity function of event time is the product of two components, $H \equiv (e^{g_{k',k}})_{K \times K}$ and ξ. The former, H, is at population level, which captures the examinee’s overall response speed, while the latter, ξ, is at individual level, which captures speed heterogeneity among different examinees. In the event history analysis literature (Allison, 1984; Hougaard, 1995; Yamaguchi, 1991), H is interpreted as a fixed effect and ξ is interpreted as a random effect (frailty). In our model, a ‘topic’ can be viewed as a class of event types sharing with the similar particular meanings. Different events, containing various meanings, may belong to distinct topics. Therefore,the topic sequence $\mathbf{z}_{1:N}$ characterizes the observed event process.

Our model connects the observed data to the latent variables. This is in the spirit of the classical item response theory (IRT) models (Embretson & Reise, 2013) and diagnostic classification models (DCMs; Rupp, Templin, & Henson, 2010). In IRT, the examinee’s ability is measured by assuming a low-dimensional model structure. The proposed model is also formulated by using the dimension reduction technique. In DCM, the Q-matrix specifies the relationship between items and latent attributes. In our model, matrix B plays a similar role. It quantifies the relationship between event types and latent topics. On the other hand, the proposed model also has its own distinct features. It uses time-stamped event process as the responses, which are no longer binary/multi-categorical. For each examinee, the sequence of latent topics can be viewed as his/her latent state. Note that, unlike in IRT/DCM, the length of the sequence is not fixed but depends on the number of actions the examinee takes.

2.3. Likelihood function

By equation (11), the likelihood function with m examinees can be written as

$$l(B,G,\mathbf{p}^0,R,a,d|\{\mathbf{e}_{1:N},\mathbf{t}_{1:N}\}_{i=1}^{m}) \\ =\prod_{i=1}^{m}\left\{\int_{\Lambda}\int_{\xi}\{\sum_{\mathbf{z}_{1:N}}\prod_{n=1}^{N}p(e_n|z_n,B)p(t_n|t_{n-1},z_{n-1},z_n,\xi,G)p(z_n|z_{n-1},\Lambda)\}p(\xi|a,d)p(\Lambda|R)\mathrm{d}\xi\mathrm{d}\Lambda\right\}. \tag{12}$$

In principle, one can get the maximum likelihood estimator (MLE) by maximizing (12). A standard approach is the EM algorithm (Bailey & Elkan, 1994; Dempster, Laird, & Rubin, 1977; Friedman, Hastie, & Tibshirani, 2001). However, in practice, it is prohibitively difficult to solve the MLE. For this particular case, it is extremely challenging to compute the posterior of latent variables, that is,

$$p(\mathbf{z},\Lambda,\xi|e,t,B,G,\mathbf{p}^0,R,a,d)=\frac{p(t,e,\mathbf{z},\Lambda,\xi|B,G,\mathbf{p}^0,R,a,d)}{p(t,e|B,G,\mathbf{p}^0,R,a,d)}. \tag{13}$$

Specifically, to calculate the denominator of (13) requires a large number of summations which grows exponentially fast as the number of events becomes large (Blei, Ng, & Jordan, 2003).

An alternative approach is the variational Bayes (VB) method (Blei, Kucukelbir, & McAuliffe, 2017), which is a modern statistical tool to approximate difficult-to-compute probability densities (Blei *et al.*, 2003; Natesan, Nandakumar, Minka, & Rubright, 2016). In contrast to sampling from a true posterior as in the traditional Monte Carlo method, VB postulates a family of distributions, assumed to have a much simpler form by reducing the dependency structure, to approximate the true posterior. The estimators are solved by maximizing a different objective function, known as the evidence lower bound (ELBO). However, the ELBO differs from and is usually smaller than the underlying log-likelihood function. Consequently, the resulting estimation may be biased; for how good the ELBO is as a proxy in some special cases, we refer to Hall, Ormerod, and Wand (2011) for the case of Poisson mixtures and You, Ormerod, and Mueller (2014) for the Bayesian linear model.

In the following two sections, we propose an empirical Bayes-type variational inference with continuous-time hidden Markov processes for event history data. Although there is a literature on variational inference for ther hidden Markov model (Foti, Xu, Laird, & Fox, 2014; Johnson & Willsky, 2014), the existing work does not cover the current setting in which the events are observed at irregular time points. In addition, we establish the usual asymptotic properties, including consistency and normality, of the associated estimators.

3. Forward-backward variational EM algorithm

In this section we introduce a forward-backward variational EM (FB-VEM) algorithm to estimate model parameters. There are two main steps in the proposed algorithm.

1. For examinee i, we consider a variational family

$$q(\mathbf{z}_i, \Lambda_i, \xi_i) = q(\xi_i|\tilde{a}_i, \tilde{d}_i)q(\mathbf{z}_i|p_i, \kappa_i)\prod_{k=1}^{K} q_k(\lambda_i^k|\gamma_i^k), \tag{14}$$

which approximates the conditional joint distribution $p(\mathbf{z}_i, \Lambda_i, \xi_i|e, t, B, G, \mathbf{p}^0, R, a, d)$. Here, $q(\xi_i|\tilde{a}_i, \tilde{d}_i)$ is the gamma density with shape $\tilde{a}_i$ and rate $\tilde{d}_i$; $q(\mathbf{z}_i|p_i, \kappa_i)$ is the (joint) probability density function of vector $\mathbf{z}_i$ (see (23) in Appendix A); $q_k(\lambda_i^k|\gamma_i^k)$ is the density function of a K-dimensional Dirichlet distribution with parameter γ_i^k. For notational simplicity, we let $q_i = q(\mathbf{z}_i, \Lambda_i, \xi_i)$ and $q = \prod_i q_i$ throughout what follows. Therefore, q is the probability density function of $(\mathbf{z}_i, \Lambda_i, \xi_i)_{i=1}^m$ with parameters $\left(\tilde{a}_i, \tilde{d}_i, p_i, \kappa_i, \{\gamma_i^k, k = 1, \ldots, K\}\right)_{i=1}^m$.
2. We define the objective function

$$EL(q, \eta) \equiv \sum_{i=1}^{m}\{\mathbb{E}_{q_i}\log(p(\mathbf{z}_i, \Lambda_i, \xi_i, \mathbf{e}_i, \mathbf{t}_i|\eta)/q_i)\}, \tag{15}$$

where $\eta = (B, G, \mathbf{p}^0, R, a, d)$. We maximize $EL(q, \eta)$ with respect to q and η by using the coordinate ascent method. We write $EL(q, \eta)$ as EL for simplicity in the remainder of the paper.

We provide some remarks to end this section. EL is known as the evidence lower bound, which is closely related to the Kullback–Leibler (KL) distance (Blei *et al.*, 2017), that is,

$$\sum_i \log p(\mathbf{z}_i, \Lambda_i, \xi_i|\mathbf{e}_i, \mathbf{t}_i) = EL + \sum_i KL(q_i \parallel p(\mathbf{z}_i, \Lambda_i, \xi_i|\mathbf{e}_i, \mathbf{t}_i, \eta)). \tag{16}$$

In other words, the log marginal likelihood equals the sum of EL and the KL distance between q and the true posterior. Therefore, among all the distributions in the variational family, a good approximation, q, should be close to the true posterior distribution in terms of KL distance. The computation is similar to that of the EM algorithm, that is, the model parameters are estimated by alternately solving the E-step and M-step. The only difference is that (16) only requires integration with respect to approximate distribution q. For our choice of variational family, each update has closed form except for R. The computation becomes much simpler as a result. The complete FB-VEM algorithm is presented in Algorithm 1, and the detailed calculations are given in Appendices A–C.

4. Theoretical properties of FB-VEM algorithm

In this section we establish some theoretical results for the FB-VEM algorithm and for parameter estimation. Specifically, we show the convergence to the locally optimal solution of our algorithm in Theorem 1 and establish the consistency and asymptotic normality of the estimators in Theorems 2–4.

Recall that the proposed estimator is the maximizer of the optimization problem

$$(\hat{q}, \hat{\eta}) = \arg\max_{q,\eta} \sum_{i=1}^{m} \left\{ \mathbb{E}_{q_i} \log(p(\mathbf{z}_i, \Lambda_i, \xi_i, \mathbf{e}_i, \mathbf{t}_i | \eta)/q_i) \right\}. \tag{17}$$

Since we are only interested in the estimation of B, G and q, we can assume the priors of Λ and ξ to be fixed without loss of generality. With a slight abuse of notation, we let $\eta = (B, G)$ be the parameter of interest and $\eta^* = (B^*, G^*)$ be the true parameter. Furthermore, we assume that the termination time τ is the same for all examinees. We denote the ELBO by EL_τ, which depends on τ implicitly.

Theorem 1 gives the local convergence of the FB-VEM algorithm. As a consequence, the proposed estimator will converge to the optimal solution when EL only admits one local maximizer or the starting point is chosen in the neighbourhood of the optimum.

Theorem 1. *The FB-VEM algorithm returns a local optimum given by (17).*

The objective function is not the log-likelihood but the evidence lower bound. The evidence lower bound is always smaller than the usual log-likelihood. Therefore, we wish to know whether or not we can consistently estimate the model parameters, including topic-word parameters (i.e., B) and topic-transition intensity parameters (i.e., G); and whether or not we can consistently estimate personal transition probabilities. Our results are stated for two situations: (1) the duration τ is bounded; (2) the duration τ goes to infinity. For (1), we show in Theorem 2 that the estimator will converge, but the limit may be different from the true parameter. For (2), we show that the estimator converges to the true parameter. Furthermore, under certain regularity conditions, the personal-specific transition probabilities can be consistently estimated when τ goes to infinity. These results are stated in Theorems 3 and 4.

Theorem 2. *Under Assumptions A1–A3 given in Appendix D, there exists a consistent estimator* $\hat{\eta}$ *such that* $\sqrt{m}(\hat{\eta} - \check{\eta}(\tau)) \rightsquigarrow N\left(0, A_1^{-1}(\tau) A_2(\tau) A_1^{-1}(\tau)\right)$.

Theorem 2 says that the proposed estimator converges to some limit $\check{\eta}(\tau)$ when the time duration τ is bounded. For each fixed τ, this may be viewed as an estimation problem under a misspecified model, as the estimating equation is constructed via the ELBO instead of log-likelihood. As a result, $\check{\eta}(\tau)$ may be different from true parameter η^*, that is, the estimator is biased when individuals are only observed for a short time.

However, when individuals are observed for a long time, we can accurately estimate the unobserved personal effect since the measurement and approximation errors will vanish. In that case, we can get consistent estimates of model parameters. The following results hold when the sample size is large and the observation time is long.

Theorem 3. *Suppose that Assumptions A1, A2, A3′ and A4′ hold and that $H_a(\eta)$ admits a unique global maximizer. Then, for any $\delta > 0$, we have that* $P\left(\hat{B} \notin B(B^*, \delta)\right) \to 0$,

$P(\hat{G} \notin B(G^*, \delta)) \to 0$, $\hat{q}(\Lambda_i \in B(\Lambda_i^*, \delta)) \to_{a.s.} 1$ and $\hat{q}(\xi_i \in B(\xi_i^*, \delta)) \to_{a.s.} 1$ for all i when $m, \tau \to \infty$.

Theorem 3 implies that the ELBO approaches the log marginal likelihood under a doubly asymptotic regime, that is, the sample size, m, is large and the observation time, τ, is long. Furthermore, we can show that the difference between *EL* and the log marginal likelihood is of order $O(1/\sqrt{\tau})$; see the Appendix S1. Therefore, we can estimate personal effects and the consistency of topic parameters follows as well.

Theorem 4. *Under Assumptions A1, A2, A3′ and A4′ and $m = O(\tau^{\delta})(\delta < 1)$, we have*

$$\sqrt{m}(\hat{\eta} - \eta^*) \rightsquigarrow N(0, Q^{-1}) \quad \text{as}\, m \to \infty. \tag{18}$$

One immediate result of Theorem 4 is that the bias of the proposed estimators is of order $o(1/\sqrt{m})$, therefore negligible when $m = \tau^{\delta}(\delta < 1)$ and $\tau \to \infty$. Proofs of Theorems 1–4 are provided in the Appendix S1.

5. Simulation study

We conducted multiple simulations, three of which are reported here, to assess the performance of the proposed estimators. Study 1 emphasizes on the mechanisms of transition structure in the proposed model. Study 2 shows the performance of the proposed method in the classical setting with moderate number of event types. Study 3 evaluates our method in a large-scale setting. The simulation results show that the proposed method works well and agrees with theoretical findings.

5.1. Study 1

This study considers the situation in which only the sequence of events is used and the time-stamps are ignored. It illustrates how the event patterns are captured by the proposed LTA model.

Table 2. Probabilities of event patterns, study 1

Event pattern	*AB*	*AD*	*CB*	*CD*	*E*	*T*
Probability	8/40	8/40	8/40	8/40	7/40	1/40

Table 3. Transition probabilities of events, study 1

	A	*B*	*C*	*D*	*E*	*T*
A	0	.5	0	.5	0	0
B	.4	0	.4	0	.175	.025
C	0	.5	0	.5	0	0
D	.4	0	.4	0	.175	.025
E	.4	0	.4	0	.175	.025
T	.4	0	.4	0	.175	.025

Table 4. Expected and estimated B and R for $K = 2$, study 1

	A	B	C	D	E	T
B						
1	0.00	0.485	0.00	0.485	0.03	0.00
2	0.41	0.00	0.41	0.00	0.15	0.03
$\hat{B}$						
1	0.00	0.485	0.00	0.485	0.03	0.00
2	0.41	0.00	0.41	0.00	0.15	0.03

	1	2
norm(R)		
1	0.00	1.00
2	0.87	0.13
norm($\hat{R}$)		
1	0.00	1.00
2	0.87	0.13

Our set up contains six different event types, A, B, C, D, E and T. Here, event T denotes termination, which is always the last event in the process. We assume that A and C can only be followed by B and D with the same probabilities, while all four share the same frequency. We sample six event patterns according to the multinomial distribution shown in Table 2 until the termination event T is sampled. The corresponding transition probabilities are provided in Table 3. We generate 100 independent copies of such event processes. Under this specification, we expect that A and C should be in the same topic; B and D should be clustered together. This is because A and B are the counterparts of C and D.

The simulated data are fitted by setting topic number $K = 2$ and $K = 3$, respectively. The parameter estimates are given in Tables 4 and 5. Here we use norm $(\hat{R}) = \left(\text{norm}\left(\hat{r}_k^{k'}\right)\right)_{k',k}$ to denote the row-normalized matrix of $\hat{R}$, where $\text{norm}\left(\hat{r}_k^{k'}\right) = \hat{r}_k^{k'} / \sum_{k=1}^{K} \hat{r}_k^{k'}$. From the two tables, we can see that the proposed method perfectly classifies the six event types into the topics as expected. Event types A and C are in the same topic, while B and D are in the other topic. Note that such clustering can not be obtained if we ignore the event transition information.

5.2. Study 2

We consider $m = 1{,}000$ users, $K = 4$ latent topics and $V = 10$ event types in this study. The topic-event matrix $B_{K \times V}$ is constructed in Table 6, where we highlight the top events in bold for every topic. We let $a = d = 1$ such that the average random effect of response time is 1. We set the initial probability of topics to be uniform, $\mathbf{p}^0 = (1/K, \ldots, 1/K)$. The parameters G and hyperparameter R are given in Table 7. For each user, we simulate the event process according to the the initial probability $\mathbf{p}^0$, the topic-event matrix $B_{K \times V}$ and the intensity parameter G until the tenth event type occurs. In this setting, users would have 500 (=1/0.002) events on average in their processes.

We simulate 100 data sets and run the routine 20 times with different initial values for each set. There are in total 5×10^5 events on average in each data set.[1] Each iteration of

[1] In this paper the computation times reported are those obtained on a PC with 2.7 GHz Intel® Core i5 processor.

Table 5. Expected and estimated B and R for $K = 3$, study 1

	A	B	C	D	E	T
B						
1	0.00	0.50	0.00	0.50	0.00	0.00
2	0.50	0.00	0.50	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.87	0.13
$\hat{B}$						
1	0.00	0.51	0.00	0.49	0.00	0.00
2	0.49	0.00	0.51	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.87	0.13

	1	2	3
norm(R)			
1	0.00	0.80	0.20
2	1.00	0.00	0.00
3	0.00	0.80	0.20
norm($\hat{R}$)			
1	0.00	0.80	0.20
2	1.00	0.00	0.00
3	0.00	0.80	0.20

Table 6. True B, study 2

	k									
v	1	2	3	4	5	6	7	8	9	10
B										
1	**0.30**	**0.30**	**0.10**	**0.10**	0.05	0.05	0.05	0.024	0.024	0.002
2	**0.10**	**0.10**	**0.30**	**0.30**	0.05	0.05	0.05	0.024	0.024	0.002
3	**0.10**	**0.10**	0.05	0.05	**0.30**	**0.30**	0.05	0.024	0.024	0.002
4	**0.10**	**0.10**	0.05	0.05	0.05	0.024	**0.30**	**0.30**	0.024	0.002

Note. The top events are highlighted in bold for every topic.

Table 7. True G and R, study 2

	k'			
k	1	2	3	4
G				
1	2	1	−1	−2
2	1	2	1	−1
3	−1	1	2	1
4	−2	−1	1	2
R				
1	40	20	5	1
2	1	40	20	5
3	5	1	40	20
4	20	5	1	40

Table 8. Estimated $\hat{B}$ and RMSE ($\times 10^2$), study 2

	k									
v	1	2	3	4	5	6	7	8	9	10
1	**0.30** (2.1)	**0.30** (2.2)	0.10 (1.2)	0.10 (1.2)	0.053 (2.1)	0.053 (2.0)	0.052 (1.5)	0.026 (1.6)	0.024 (0.089)	0.0021 (0.047)
2	**0.10** (2.1)	**0.10** (2.1)	**0.30** (2.0)	**0.30** (1.9)	0.049 (0.27)	0.049 (0.26)	0.050 (0.11)	0.024 (0.12)	0.024 (0.065)	0.0020 (0.025)
3	**0.10** (0.34)	**0.10** (0.35)	0.050 (1.2)	0.050 (1.2)	**0.30** (1.0)	**0.30** (1.1)	0.049 (0.28)	0.023 (0.32)	0.024 (0.077)	0.0020 (0.032)
4	**0.10** (0.12)	**0.10** (0.12)	0.050 (0.14)	0.050 (0.14)	0.050 (0.21)	0.024 (0.17)	**0.30** (0.27)	**0.30** (0.28)	0.024 (0.067)	0.0020 (0.021)

Note. The top events are highlighted in bold for every topic.

Table 9. Estimated $\hat{G}$ and RMSE, study 2

	k'			
k	1	2	3	4
1	1.9 (0.59)	0.97 (0.34)	−0.93 (0.37)	−1.9 (0.56)
2	1.1 (0.48)	2.0 (0.13)	1.0 (0.16)	−1.0 (0.33)
3	−0.96 (0.33)	1.0 (0.46)	2.0 (0.12)	0.99 (0.14)
4	−1.9 (0.47)	−1.0 (0.25)	0.97 (0.74)	2.0 (0.11)

Table 10. True R, estimated $\hat{R}$ after normalization and RMSE, study 2

	k'			
k	1	2	3	4
norm(R)				
1	0.606	0.303	0.076	0.015
2	0.015	0.606	0.303	0.076
3	0.076	0.015	0.606	0.303
4	0.303	0.076	0.015	0.606
norm($\hat{R}$)				
1	0.595 (0.081)	0.300 (0.017)	0.080 (0.029)	0.024 (0.048)
2	0.023 (0.010)	0.608 (0.026)	0.294 (0.022)	0.075 (0.013)
3	0.074 (0.014)	0.022 (0.014)	0.609 (0.016)	0.295 (0.017)
4	0.289 (0.048)	0.083 (0.048)	0.016 (0.004)	0.612 (0.007)

the FB-VEM algorithm takes about 3 s on average, and the whole estimation procedure completes within 250 iterations. The final estimates $\hat{B}$, $\hat{G}$, normalized $\hat{R}$ and their RMSEs are given by Tables 8–10. From the estimators, we can see that our model successfully captures most of the signals, that is, the estimated parameters are very close to the truth. As we have mentioned, the (k', k)th entry of the normalized R is the expectation of topic assignment parameter $\lambda_k^{k'}$. We focus more on the normalized version instead of $\hat{R}$ itself because it gives the probabilities of topic transitions and is more closely related to behaviour patterns.

5.3. Study 3

In this study the performance of our model is evaluated for large data sets. We consider $m = 5{,}000$ users, $K = 8$ latent topics and $V = 1{,}000$ event types. We set the kth row of B, $1 \leq k \leq K$, as in Table 11. We repeat the same procedure as described in study 2 and get the estimated results. Here each simulated data set contains about 5×10^5 events. It takes around 7.5 s to finish one iteration on average, and the whole estimation completes within 300 iterations.

In this study, one way to evaluate the performance of our model is to see whether we can identify the top events with large probabilities and prevent the events with small probabilities from popping up to the top list. Following this idea, we use a cut-off point b_0 to divide all events into two groups. Let $c_{k,v} \in \{1, 2\}$ denote the true membership of the vth event in topic k, then let.

Table 11. True B, study 3

v	1	…	$9(k-1)$	$9(k-1)+1$	$9(k-1)+2$	$9(k-1)+3$	$9(k-1)+4$
$b_{k,v}$	5×10^{-4}	…	5×10^{-4}	0.3	0.1	0.05	0.02
v	$9(k-1)+5$	$9(k-1)+6$	$9(k-1)+7$	$9(k-1)+8$	…	999	1,000
$b_{k,v}$	0.02	0.003	0.001	5×10^{-4}	…	5×10^{-4}	0.01

$$c_{k,v}=\begin{cases}1, & b_{k,v}\geq b_0,\\ 2, & b_{k,v}\leq b_0.\end{cases}\tag{19}$$

The estimated membership $\hat{c}_{k,v}$ is defined in a similar way, equal to 1 if $\hat{B}_{k,v}\geq b_0$ and 2 otherwise. We introduce an index

$$CR=\frac{1}{K\cdot V}\sum_{k,v}\mathrm{I}\{c_{k,v}=\hat{c}_{k,v}\},\tag{20}$$

which takes values from [0, 1]. This index measures the consistency of the memberships; that is, a larger CR implies a better model fit. We let b_0 = 0.005, 0.015, 0.025, 0.075, 0.15 and find that it is more challenging to estimate $c_{k,v}$ when b_0 = 0.025. In other words, CR achieves its minimum value at b_0 = 0.025. The average of CR for 100 sets of simulations equals 99.89%. The result suggests that top events in topics can be successfully detected.

6. Application to climate control data

We apply our method to the climate control item in PISA 2012 as described in Section 2. The log file of this item contains individual event process histories. The data set we use here includes 16,920 students, 54.4% of whom answered the item correctly. On average, it takes around 9 actions for a student to explore the item (excluding drawing lines in the diagram), requiring about 2 min. We remove the 'START_ITEM', 'END_ITEM' and all 'Diagram' events. Then we use a three-dimensional vector to denote the remaining 'apply' events, with each entry taking a value from $\{-2,-1,0,1,2\}$. The value here represents the position of the corresponding control slider. For instance, if a student moves the top control to '2' while keeping the other two controls at '▲' and then clicks on the 'apply' button (see the fourth event, e_4, in Table 1), then the event is coded as (2, 0, 0).

We fit the model with a series of topic numbers. It turns out that the events with top probabilities are similar across the topics when $K>4$. Therefore, the parameter estimates we present here are the results when $K=4$. The climate control data set contains around 5.3×10^4 events in total. When $K=4$, each iteration of the FB-VEM algorithm takes about 1 s on average. It takes less than 600 iterations to finish the whole estimation procedure. To compute the standard errors of the estimated parameters, we use the parametric bootstrap method by simulating 100 data sets based on the estimated model. For each data set generated, we apply the FB-VEM algorithm to obtain the corresponding parameter estimates. We report the standard errors by calculating the standard deviations of 100 sets of estimates.

Table 12 shows the four topics with their top events, and the initial topic distribution is given in Table 13. We can see that both topics 1 and 2 contain event types with at most one moved control at a time, which are the most efficient ways to explore each control. Most examinees will start with events in those two topics according to Table 13. Apart from the

Table 12. The four topics with their top events, Programme for International Student Assessment (PISA) 2012 climate control item

Topic	Top events						
1	(1, 0, 0)	(0, 0, 0)	Reset				
2	(0, 1, 0)	(0, 2, 0)	(0, 0, 1)	(0, 0, 2)	(0, 0, 0)	(2, 0, 0)	(1, 0, 0)
3	(2, 2, 2)	(−2, −2, −2)	(1, 1, 1)				
4	Reset						

Table 13. Estimated $\mathbf{p}^0$ and the standard errors ($\times 10^3$), Programme for International Student Assessment (PISA) 2012 climate control item

1	2	3	4
0.66 (4.7)	0.28 (4.2)	0.06 (2.7)	0.00 (1.1)

Table 14. Estimated G and standard errors ($\times 10^2$), Programme for International Student Assessment (PISA) 2012 climate control item

	k'			
k	1	2	3	4
1	−2.64 (1.1)	−2.69 (2.0)	−2.60 (2.4)	−2.58 (2.4)
2	−230.26 (1.0)	−1.82 (0.64)	−2.03 (1.8)	−1.76 (0.62)
3	−2.03 (7.5)	−2.18 (3.4)	−1.86 (0.73)	−2.17 (2.1)
4	−3.31 (4.7)	−1.58 (0.67)	−2.89 (8.7)	−2.59 (10)

top events, topic 3 includes almost all events with more than one moved control, and their probabilities within topic 3 are quite even. 'Reset' seems crucial to this item since it is dominant in topic 4, though its position in the processes could be different. The estimated $G = (g_{k',k})_{k',k}$ in Table 14 indicates how fast the examinees would have events from one topic to another. It seems that topic 2 often comes right after topic 4, and almost no one would jump to topic 1 once they have some events from topic 2. We also find that it usually takes a shorter time to have event types within the same topic.

We further analyse different behavioural patterns of examinees by looking at their person-specific parameters. Here for student i, we use the posterior mean of topic assignment parameter, λ_i^k, as the individual transition probabilities, which could be approximated by the normalized γ_i^k. We denote this by norm(γ_i^k). It not only contains

Table 15. *k*-means results, Programme for International Student Assessment (PISA) 2012 climate control item

Cluster	1	2	3	4
Cluster size	4,490	3,706	7,187	1,537
Correct rate	81.5%	73.4%	37.0%	11.0%

Table 16. *k*-means centers of clusters 1 and 4, Programme for International Student Assessment (PISA) 2012 climate control item

	1	2	3	4
Cluster 1				
1	0.49	0.20	0.16	0.15
2	0.00	0.20	0.05	0.75
3	0.03	0.06	0.80	0.11
4	0.04	0.95	0.01	0.01
Cluster 4				
1	0.31	0.19	0.30	0.20
2	0.00	0.40	0.36	0.24
3	0.32	0.08	0.35	0.25
4	0.20	0.33	0.15	0.32

information about the topic transition patterns among the whole population, but also captures the personal-level variation. We then apply the k-means method to $\{\text{norm}(\gamma_i^k), k = 1,\ldots, K\}$. As shown in Table 15, it turns out that the result is meaningful when the total population is divided into four clusters. According to the average correct rate, the topic transitions do contain significant information about the item.

We also present the centres of clusters 1 and 4 in Table 16 since their average correct rates differ widely. Transition probabilities between topics 3 and 4 differ substantially across the clusters. These two transition matrices also reveal the learning trajectories of examinees. Topic 1 (see Table 13) is the dominant initial topic. It is mainly about the top control. After the first attempt, around half of the students in cluster 4 would move on to the rest of the controls and attempt to move multiple bars at the same time (transit from topic 1 to topics 2 and 3). They are more likely to keep learning without using 'reset' (stay in topic 2 or 3). Notice that the central and bottom controls are both about humidity; moving more than one slider at a time could lead to confusion. That might be the reason for their low correct rate.

Students in cluster 1, after exploring the top control (topic 1), tend to either click on 'reset' (transit from topic 1 to topic 4 or stay at topic 1) and then start to move the second or third control (transit from topic 1 or 4 to topic 2), or just go on without clearing the panel (transit from topic 1 to topic 2). Once they reach topic 2, they could explore the second control, click on 'reset' to clear the panel and then try to solve the last control (transit to topic 4 and then go back to topic 2). The main strategy behind this systematic behaviour path is divide and conquer.

7. Conclusion

In this paper we propose a latent topic model to analyse process data. Based on a hierarchical Bayesian continuous-time model, we add a hidden Markov structure. We apply the proposed method to the climate control item in PISA 2012. The proposed model clusters the event types into four latent topics to capture the key features of the test item. Based on the topic transitions of each examinee, we further classify the population into four groups and investigate the learning trajectories. It turns out that the strategy known as divide and conquer plays an essential role in solving the item.

The latent topic model with the proposed FB-VEM algorithm is a general method that could be applied to other CPS items and other kinds of process data such as log files recorded on websites. Once the event type is properly defined, the behaviour patterns could be learned through topics and their transitions. Though our approach could be used as a first step to understand the process data, certain domain knowledge is still required to interpret each topic, as with most unsupervised methods.

The proposed approach may be extended to include baseline covariates such as gender and nationality. The latent Markov structure may also be extended so that the current state is related to the entire past history. As far as computation is concerned, since the event processes in the FB-VEM algorithm share only a few common parameters, most user-level parameters could be updated separately at each iteration. Consequently, the distributed algorithm may reduce the computational burden. Currently, there is no effective method to compute the standard errors of variational Bayes estimators, which is an important problem for further investigation.

The authors thank the Editor, the Associate Editor,and the reviewers for many helpful and constructive comments. The research is supported by US National Science Foundation SES-1826540, IIS-1633360.

Conflict of interest

All authors declare no conflict of interest.

Author contributions

Haochen Xu (Conceptualization; Data curation; Formal analysis; Methodology; Software; Validation; Visualization; Writing – original draft; Writing – review & editing) Guanhua Fang (Conceptualization; Data curation; Formal analysis; Methodology; Software; Validation; Visualization; Writing – original draft; Writing – review & editing) Zhiliang Ying (Conceptualization; Funding acquisition; Methodology; Project administration; Resources; Supervision; Writing – original draft; Writing – review & editing).

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data*. Beverly Hills, CA: Sage. https://doi.org/10.4135/9781848608184.n16

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers. In *Proceedings of International Conference on Intelligent Systems for Molecular Biology* (Vol. *2*, pp. 28–36). Menlo Park, CA: AAAI Press.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*, 859–877. https://doi.org/10.1080/01621459.2017.1285773

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). Statistical analysis of complex problem-solving process data: An event history analysis approach. *Frontiers in Psychology*, *10*, 486. https://doi.org/10.3389/fpsyg.2019.00486

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1–38. https://doi.org/10.1142/97898123887590028

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Hove, UK: Psychology Press. https://doi.org/10.1007/springerreference_183948

Fischer, A., Greiff, S., & Funke, J. (2011). The process of solving complex problems. *Journal of Problem Solving*, *4*(1), 19–42. https://doi.org/10.7771/1932-6246.1118

Foti, N., Xu, J., Laird, D., & Fox, E. (2014). Stochastic variational inference for hidden Markov models. In M. I. Jordan, Y. LeCun & S. A. Solla (Eds.), *Advances in neural information processing systems* (pp. 3599–3607). Red Hook, NY: Curran.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. New York, NY: Springer. https://doi.org/10.1007/BF02985802

Goodman, M., Finnegan, R., Mohadjer, L., Krenzke, T., & Hogan, J. (2013). *Literacy, numeracy, and problem solving in technology-rich environments among us adults: Results from the Program for the International Assessment of Adult Competencies 2012. First look (NCES 2014–008)*. ERIC. https://doi.org/10.1787/9789264128859-en

Griffin, P., McGaw, B., & Care, E. (2012). *Assessment and teaching of 21st century skills*. Dordrecht, The Netherlands: Springer. https://doi.org/10.1007/978-94-007-2324-5_5

Hall, P., Ormerod, J. T., & Wand, M. P. (2011). Theory of Gaussian variational approximation for a Poisson mixed model. *Statistica Sinica*, *21*, 369–389. https://doi.org/10.1198/jcgs.2011.09118

Halpin, P. F., & De Boeck, P. (2013). Modelling dyadic interaction with Hawkes processes. *Psychometrika*, *78*, 793–814. https://doi.org/10.1007/s11336-013-9329-1

He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with *n*-grams. *Quantitative psychology research* (pp. 173–190). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-19977-1_13

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with *n*-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 750–777). Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-4666-9441-5.ch029

He, Q., von Davier, M., Greiff, S., Steinhauer, E. W., & Borysewicz, P. B. (2017). Collaborative problem solving measures in the Programme for International Student Assessment (PISA). In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 95–111). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-33261-1_7

Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis*, *1*, 255–273.

Johnson, M., & Willsky, A. (2014). Stochastic variational inference for Bayesian time series models. In *Proceedings of the 31st International Conference on Machine Learning* (pp. 1854–1862).

Natesan, P., Nandakumar, R., Minka, T., & Rubright, J. D. (2016). Bayesian prior choice in IRT estimation using MCMC and variational BAYES. *Frontiers in Psychology*, 7, 1422. https://doi.org/10.3389/fpsyg.2016.01422

OECD (2014). *PISA 2012 technical report*. Paris, France: OECD Publishing. Retrieved from http://www.oecd.org/pisa/pisaproducts/pisa2012technicalreport.html

OECD (2016). *PISA 2015 results in focus*. Paris, France: OECD Publishing. Retrieved from https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

Polyak, S. T., von Davier, A. A., & Peterschmidt, K. (2017). Computational psychometrics for the measurement of collaborative problem solving skills. *Frontiers in Psychology*, *8*, 2029. https://doi.org/10.3389/fpsyg.2017.02029

Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: A didactic. *Frontiers in Psychology*, *9*, 2231. https://doi.org/10.3389/fpsyg.2018.02231

Rabiner, L. R., & Juang, B.-H. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, *3*(1), 4–16.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press. https://doi.org/10.1111/j.1745-3984.2011.00141.x

Vista, A., Care, E., & Awwal, N. (2017). Visualising and examining sequential actions as behavioural paths that can be interpreted as markers of complex behaviours. *Computers in Human Behavior*, *76*, 656–671. https://doi.org/10.1016/j.chb.2017.01.027

Xu, H., Fang, G., Chen, Y., Liu, J., & Ying, Z. (2018). Latent class analysis of recurrent events in problem-solving items. *Applied Psychological Measurement*, *42*, 478–498. https://doi.org/10.1177/0146621617748325

Yamaguchi, K. (1991). *Event history analysis*. Newbury Park, CA: Sage.

You, C., Ormerod, J. T., & Mueller, S. (2014). On variational Bayes estimation and variational information criteria for linear regression models. *Australian and New Zealand Journal of Statistics*, *56*(1), 73–87. https://doi.org/10.1111/anzs.12063

Received 23 July 2019; revised version received 25 October 2019

Supporting Information

The following supporting information may be found in the online edition of the article:

Appendix S1. Technical proofs and additional results.

Appendix A:

FB-VEM algorithm

In this section we provide a detailed description of the FB-VEM algorithm. To find a suitable posterior $q(\mathbf{z}_i, \Lambda_i, \xi_i)$ which is simple enough and could approximate the true posterior well (equation (13)), that is,

$$q(\mathbf{z}_i, \Lambda_i, \xi_i) \approx p(\mathbf{z}_i, \Lambda_i, \xi_i | e, t, B, G, \mathbf{p}^0, R, a, d),$$

we construct posterior $q\,(\cdot)$ for Λ_i, $\mathbf{z}_i$, and ξ_i separately.

For Λ_i, we choose a distribution from the variational family

$$q(\Lambda_i) = \prod_{k=1}^{K} q_k\left(\lambda_i^k | \gamma_i^k\right). \tag{21}$$

$q_k(\lambda_i^k|\gamma_i^k)$ is set as a K-dimensional Dirichlet with parameters $\gamma_i^j = \left(\gamma_{i,1}^k, \ldots, \gamma_{i,K}^k\right)$, since the exact conditional distribution of λ_i^k is Dirichlet,

$$p(\lambda_i^k | z, R) = \mathrm{Dir}_K\left(\mathbf{r}^k + \sum_{n=1}^{N_i} \mathrm{I}\{z_{i,n} = k\} \cdot \mathbf{z}_{i,n+1}\right), \quad 1 \le k \le K. \tag{22}$$

Here, $\mathbf{z}_{i,n+1}$ is a K-vector where the $z_{i,n+1}$th element is 1 and all others are equal to 0. For $\mathbf{z}_i$, we let it follow a multinomial such that

$$\begin{aligned} q(\mathbf{z}_i | p_i, \kappa_i) \propto & \left(\prod_{n=2}^{N_i} \kappa_i e^{g_{z_{i,n-1} z_{i,n}}} \exp\left\{-\left(t_{i,n} - t_{i,n-1}\right) \kappa_i e^{g_{z_{i,n-1} z_{i,n}}}\right\} p\left(e_{i,n} | z_{i,n}, B\right) p_{i,z_{i,n}}^{z_{i,n}}\right) \\ & \cdot p(e_{i,1} | z_{i,1}, B) p(z_{i,1} | \mathbf{p}^0). \end{aligned} \tag{23}$$

Finally, we set $q(\xi_i | \tilde{a}_i, \tilde{d}_i)$ to be a gamma distribution with parameter $\tilde{a}_i$ and $\tilde{d}_i$, the same as its exact posterior

$$p(\xi_i|t,z,G,a,d) = \text{Gamma}\left(N_i + a - 1, d + \sum_{n=2}^{N_i} e^{g_{z_{i,n-1},z_{i,n}}}\left(t_{i,n} - t_{i,n-1}\right)\right). \tag{24}$$

After specifying each component, the approximate posterior $q(\mathbf{z}_i, \Lambda_i, \xi_i)$ has the form

$$q(\mathbf{z}_i, \Lambda_i, \xi_i) = q(\xi_i|\tilde{a}_i, \tilde{d}_i)q(\mathbf{z}_i|p_i, \kappa_i)\prod_{k=1}^{K} q_k(\lambda_i^k|\gamma_i^k). \tag{25}$$

In order to get the optimal $q(\mathbf{z}_i, \Lambda_i, \xi_i)$, we use the Kullback–Leibler divergence as the measure to quantify its distance to $p(\mathbf{z}_i, \Lambda_i|\cdot)$, that is,

$$\mathrm{KL}(p_i, \gamma_i) = \mathrm{KL}(q(\mathbf{z}_i, \Lambda_i)||p(\mathbf{z}_i, \Lambda_i|\cdot)). \tag{26}$$

The corresponding optimizers are

$$\gamma_i^k = \mathbf{r}^k + \sum_{n=1}^{N_i}\left(\tilde{\phi}_{i,n}^{(k,1)}, \ldots, \tilde{\phi}_{i,n}^{(k,K)}\right)^{\mathrm{T}}, \tag{27}$$

$$p_{i,k}^{k'} \propto \exp\left\{\mathbb{E}_{q(\Lambda_i)}[\log \Lambda_{i,k}^{k'}]\right\} = \exp\left\{\Psi\left(\gamma_{i,k}^{k'}\right) - \Psi\left(\sum_{s=1}^{K}\gamma_{i,s}^{k'}\right)\right\}, \tag{28}$$

$$\tilde{a}_i = N_i + a - 1, \tag{29}$$

$$\tilde{d}_i = d + \sum_{n=2}^{N_i}\left(\sum_{k,l}\tilde{\phi}_{i,n-1}^{(k,l)} e^{g_{k,l}}\right)\left(t_{i,n} - t_{i,n-1}\right), \tag{30}$$

$$\kappa_i = \tilde{a}_i/\tilde{d}_i. \tag{31}$$

One of the key elements in the equations above, $\tilde{\phi}_{i,n}$, is given in equation (35).

Turning to the approximate marginal posterior of $z_{i,n}$, we calculate this by using the forward-backward algorithm. We let forward and backward functions for subject i at time $t_{i,n}$ be $f_{i,n} = (f_{i,n}(1), \ldots, f_{i,n}(K))$ and $b_{i,n} = (b_{i,n}(1), \ldots, b_{i,n}(K))$, respectively; and we let $\phi_{i,n} = \left(\phi_{i,n}^{(1)}, \ldots, \phi_{i,n}^{(K)}\right)$, $\tilde{\phi}_{i,n} = \left(\phi_{i,n}^{(k',k)}\right)_{k',k}$ be the approximate marginal posteriors of $z_{i,n}$ and $(z_{i,n}, z_{i,n+1})$, respectively, where

$$\phi_{i,n}^{(k)} = p(z_{i,n} = k|t_i, \mathbf{e}_i, B, G, \mathbf{p}^0, R), \tag{32}$$

$$\tilde{\phi}_{i,n}^{(k',k)} = p(z_{i,n} = k', z_{i,n+1} = k|t_i, \mathbf{e}_i, B, G, \mathbf{p}^0, R). \tag{33}$$

The posteriors then satisfy

$$\phi_{i,n}^{(k)} \propto f_{i,n}(k) \cdot b_{i,n}(k), \tag{34}$$

$$\tilde{\phi}_{i,n}^{(k',k)} \propto f_{i,n}(k') \cdot b_{i,n+1}(k) \cdot \tilde{p}_{i,k}^{k'} \cdot \exp\left(g_{k',k} - \kappa_i e^{g_{k',k}}\left(t_{i,n+1} - t_{i,n}\right)\right) \cdot b_{k,e_{i,n+1}}. \tag{35}$$

In the last part of the algorithm, we iteratively do the E-step and the M-step. We let $\eta = \{B, G, \mathbf{p}^0, R, a, d\}$, $\zeta = \{\phi, \tilde{\phi}, \gamma, \tilde{\mathbf{a}}, \tilde{\mathbf{d}}, p, \kappa\}$. In the E-step we calculate $Q(\eta|\zeta^{(n+1)})$ by

$$\mathbb{E}_{q(z,\theta,\xi)}[\log p(t, e, z, \theta, \xi)|\zeta^{(n+1)}], \tag{36}$$

then in the M-step we solve the optimization problem

$$\eta^{(n+1)} = \arg\max_{\eta} Q(\eta|\zeta^{(n+1)}). \tag{37}$$

As two main parts of the FB-VEM algorithm, the forward-backward algorithm and expectation-maximization algorithm are described in more detail in Appendix B and Appendix C, respectively.

Appendix B:

Forward-backward algorithm

We introduce the forward-backward algorithm and show its application in our model. The algorithm enables us to calculate the posterior distribution of latent variables (states) $\{Y_n\}$ (the latent topic $z_{i,n}$ in our model) given a series of observations $\{X_n\}$ (such as the observed event $e_{i,n}$, the event time $t_{i,n}$) in a hidden Markov model (Rabiner & Juang, 1986). Suppose there are N time-stamps in total. We let $Y_{1:n}$ and $X_{1:n}$ denote the latent variables and observations from time t_1 to t_{n}, $1 \leq n \leq N$. A key property of the hidden Markov model is $P(X_n|Y_{1:n}) = P(X_n|Y_n)$ and $P(X_n|Y_{n:N}) = P(X_n|Y_n)$, that is, the observation X_n at time t_n is independent of other latent variables once given its hidden state Y_n. Now for a specific n, we can apply the property and calculate the conditional probability as

$$P(Y_n|X_{1:N}) = P(Y_n|X_{1:n}, X_{(n+1):N}) \propto P(Y_n|X_{1:n}) \cdot P(X_{(n+1):N}|Y_n). \tag{38}$$

Here $P(Y_n|X_{1:n})$ and $P(X_{(n+1):N}|Y_n)$ are called forward probability and backward probability, which are the two major parts we are trying to obtain in our algorithm. We will derive the recursive formula for these two parts in the following subsections.

B.1. Forward probabilities and forward functions

We assume that the latent variable Y_n can take values from $\{1, \ldots, K\}$. Then the initial forward probability at time t_1 can be calculated by

$$P(Y_1 = k|X_1) \propto P(X_1, Y_1 = k) = P(X_1|Y_1 = k) \cdot P(Y_1 = k), \quad 1 \leq k \leq K, \tag{39}$$

where $P(Y1 = k)$ depends only on the initial distribution of the latent variables. We introduce the forward functions $\mathbf{f}_1 = (f_1(1), \ldots, f_1(K))$ for simplicity such that

$$f_1(k) = P(X_1|Y_1 = k) \cdot P(Y_1 = k). \tag{40}$$

At the second time-stamp t_2, we have

$$\begin{aligned} P(Y_2 = k|X_{1:2}) \propto P(X_{1:2}, Y_2 = k) &= \sum_{k'=1}^{K} P(X_{1:2}, Y_2 = k|Y_1 = k')P(Y_1 = k') \\ &= \sum_{k'=1}^{K} P(X_{1:2}|Y_2 = k, Y_1 = k')P(Y_2 = k|Y_1 = k')P(Y_1 = k'). \end{aligned}$$

Notice that

$$P(X_{1:2}|Y_2 = k, Y_1 = k') = P(X_1|Y_1 = k')P(X_2|Y_2 = k), \tag{41}$$

and we can further derive

$$\begin{aligned} P(Y_2 = k|X_{1:2}) &\propto \sum_{k'=1}^{K} P(X_2|Y_2 = k)P(Y_2 = k|Y_1 = k')P(X_1|Y_1 = k')P(Y_1 = k') \\ &= \sum_{k'=1}^{K} f_1(k) \cdot p_k^{k'} \cdot P(X_2|Y_2 = k), \end{aligned} \tag{42}$$

where $p_k^{k'} = P(Y_2 = k|Y_1 = k')$ is the transition probability from the hidden state k' to k. We continue to use $\mathbf{f}_2 = (f_2(1), \ldots, f_2(K))$ to denote

$$f_2(k) = \sum_{k'=1}^{K} f_1(k) \cdot p_k^{k'} \cdot P(X_2|Y_2 = k). \tag{43}$$

In general, given $\mathbf{f}_{n-1} = (f_{n-1}(1), \ldots, f_{n-1}(K))$ at time t_{n-1}, the forward probability at time t_n can be obtained as

$$\begin{aligned} P(Y_n = k|X_{1:n}) \propto P(X_{1:n}, Y_n = k) &= \sum_{k'=1}^{K} P(X_{1:n}, Y_n = k|Y_{n-1} = k')P(Y_{n-1} = k') \\ &= \sum_{k'=1}^{K} P(X_n|Y_n = k)P(Y_n = k|Y_{n-1} = k')P\left(X_{1:(n-1)}|Y_{n-1} = k'\right)P(Y_{n-1} = k') \\ &= \sum_{k'=1}^{K} f_{n-1}(k) \cdot p_k^{k'} \cdot P(X_n|Y_n = k). \end{aligned} \tag{44}$$

And the kth element of $\mathbf{f}_n = (f_{\mathrm{n}}(1), \ldots, f_{\mathrm{n}}(K))$ is obtained as

$$f_n(k) = \sum_{k'=1}^{K} f_{n-1}(k) \cdot p_k^{k'} \cdot P(X_n|Y_n = k). \tag{45}$$

In our model, the observations for subject i include the detailed events and the response time, so $X_{i,n} = \{e_{i,n}, t_{i,n}\}$, while the latent variable is the topic $Y_{i,n} = z_{i,n}$. A main difference between our model and the hidden Markov model in this algorithm is that the event time from $t_{i,n-1}$ to $t_{i,n}$, characterized by the intensity function, depends on both the previous and the current topic. It should be regarded as an obervation related to the topic transition. So $P(X_{1:n}, Y_n = k|Y_{n-1} = k')$ in our case should be

$$\begin{aligned} P\big(\mathbf{e}_{i,1:n}, \mathbf{t}_{i,1:n}, z_{i,n} = k|z_{i,n-1} = k'\big) = {} & P(e_{i,n}|z_{i,n} = k) \\ & \cdot P(t_{i,n}|t_{i,n-1}, z_{i,n} = k, z_{i,n-1} = k') \\ & \cdot P\big(z_{i,n} = k|z_{i,n-1} = k'\big) \\ & \cdot P(\mathbf{e}_{i,1:(n-1)}, \mathbf{t}_{i,1:(n-1)}|z_{i,n-1} = k'). \end{aligned} \tag{46}$$

Notice that $P\big(z_{i,1} = k\big) = p_k^0$ and $P(e_{i,n}|z_{i,n} = k) = b_{k,e_{i,n}}$. The forward functions of subject i, for $n = 1, \ldots, N_i$, are

$$f_{i,1}(k) = p_k^0 \cdot b_{k,e_{i,1}}, \tag{47}$$

$$f_{i,n}(k) = \sum_{k'=1}^{K} f_{i,n-1}(k) \cdot p_{i,k}^{k'} \cdot \exp\big(\lambda_{k',k} - \kappa_i e^{\lambda_{k',k}}\big(t_{i,n+1} - t_{i,n}\big)\big) \cdot b_{k,e_{i,n}}. \tag{48}$$

B.2. Backward probabilities and backward functions

The backward probabilities start from the last state of the Markov chain (at time t_N), and from there we calculate the probability at each time-stamp backwards. We assume that the initial backward function $\mathbf{b}_N = (b_N(1), \ldots, b_N(K))$ is

$$\mathbf{b}_N = (1, \ldots, 1). \tag{49}$$

This is because there are no more observations after time t_N, so we can simply set each of them to be 1. At time t_{N-1}, by definition the backward probability is

$$\begin{aligned} P(X_N|Y_{N-1} = k) &= \sum_{k'=1}^{K} P(X_N|Y_{N-1} = k, Y_N = k') \cdot P(Y_N = k'|Y_{N-1} = k) \\ &= \sum_{k'=1}^{K} b_N(k) \cdot p_{k'}^k \cdot P(X_N|Y_N = k'), \end{aligned} \tag{50}$$

where $p_{k'}^k = P(Y_N = k'|Y_{N-1} = k)$. Then we let each element of the backward function $b_{N-1} = \big(b_{N-1}(1), \ldots, b_{N-1}(K)\big)$ be

$$b_{N-1}(k) = \sum_{k'=1}^{K} b_N(k) \cdot p_{k'}^k \cdot P(X_N|Y_N = k'). \tag{51}$$

In general, given $b_{n+1} = (b_{n+1}(1), \ldots, b_{n+1}(K))$, the backward probability at time t_n is

$$\begin{aligned}
P(X_{(n+1):N}|Y_n=k) &= \sum_{k'=1}^{K} P(X_{(n+1):N}|Y_n=k, Y_{n+1}=k') \cdot P(Y_{n+1}=k'|Y_n=k) \\
&= \sum_{k'=1}^{K} P(X_{n+1}|Y_{n+1}=k') \cdot P\big(X_{(n+2):N}|Y_{n+1}=k'\big) \cdot P(Y_{n+1}=k'|Y_n=k) \\
&= \sum_{k'=1}^{K} b_{n+1}(k) \cdot p_{k'}^{k} \cdot P(X_{n+1}|Y_{n+1}=k'),
\end{aligned} \tag{52}$$

so the corresponding backward function is

$$b_n(k) = \sum_{k'=1}^{K} b_{n+1}(k) \cdot p_{k'}^{k} \cdot P(X_{n+1}|Y_{n+1} = k'). \tag{53}$$

Now we can apply the formulas to our model and get the backward functions from $n = N_i$ to $n = 1$ for each subject i as

$$b_{i,N_i}(k) = 1, \tag{54}$$

$$b_{i,n}(k) = \sum_{k'=1}^{K} b_{i,n+1}(k') \cdot p_{i,k'}^{k} \cdot \exp\big(\lambda_{k,k'} - \kappa_i e^{\lambda_{k,k'}}(t_{i,n+1} - t_{i,n})\big) \cdot b_{k',e_{i,n+1}}. \tag{55}$$

B.3. Posterior distributions of latent variables

Once we obtain the forward and backward functions, the posterior distribution of each latent variable can be calculated as

$$P(Y_n = k|X_{1:N}) \propto P(Y_n|X_{1:n}) \cdot P(X_{(n+1):N}|Y_n) \propto f_n(k) \cdot b_n(k). \tag{56}$$

The last thing we need in our algorithm is the joint posterior distribution, which is used to update other parameters in our model. It can be shown that

$$\begin{aligned}
P(Y_n = k, Y_{n+1} = l|X_{1:N}) &\propto P(X_{1:N}, Y_n = k, Y_{n+1} = l) \\
&\propto P(X_{1:n}|Y_n = k) \cdot P(X_{(n+1):N}, Y_{n+1} = l|Y_n = k) \cdot P(Y_n = k) \\
&\propto P(Y_n = k|X_{1:n}) \cdot P\big(X_{(n+2):N}|Y_{n+1} = l\big) \\
&\quad \cdot P(X_{n+1}|Y_{n+1} = l) \cdot P(Y_{n+1} = l|Y_n = k) \\
&= f_n(k) \cdot b_{n+1}(l) \cdot p_l^k \cdot P(X_{n+1}|Y_{n+1} = l).
\end{aligned} \tag{57}$$

Then the corresponding formulas for our model are

$$P(z_{i,n} = k|e_{i,1:(N_i)}, t_{i,1:(N_i)}) \propto f_{i,n}(k) \cdot b_{i,n}(k), \tag{58}$$

$$P\left(z_{i,n} = k, z_{i,n+1} = l | e_{i,1:(N_i)}, t_{i,1:(N_i)}\right) \propto f_{i,n}(k) \cdot b_{i,n+1}(l) \cdot p_{i,l}^{k} \\ \cdot \exp\left(g_{k',k} - \kappa_i e^{g_{k',k}}(t_{i,n+1} - t_{i,n})\right) \cdot b_{l,e_{i,n+1}}. \tag{59}$$

Appendix C:

Expectation-maximization algorithm

We present here the details of parameter estimation using the EM algorithm. There are two steps in the classical EM algorithm: an expectation step (E-step) and a maximization step (M-step). Given the observed data X (such as the observed event $e_{i,n}$, the event time $t_{i,n}$), the unobserved data Y (the latent topic $z_{i,n}$, the topic assignment parameter λ_i^j in our case), and a set of unknown parameters η (the topic to event probability matrix B, intensity-related matrix Λ, hyperparameter α, etc.), we define the complete-data likelihood as

$$L(\eta; X, Y) = p(X, Y|\eta),$$

and the log-likelihood as $l(\eta; X, Y) = \log L(\eta; X, Y)$.

The EM algorithm iteratively applies the two steps until convergence. In our case, given parameter values $\eta^{(n)} = \{B^{(n)}, G^{(n)}, (\mathbf{p}^0)^{(n)}, R^{(n)}\}$ obtained in the nth iteration, we first update $\zeta = \left\{\phi, \tilde{\phi}, \gamma\right\}$ using equations (34), (35) and (27) to get $\zeta^{(n+1)} = \left\{\phi^{(n+1)}, \tilde{\phi}^{(n+1)}, \gamma^{(n+1)}\right\}$ in the (n + 1)th iteration. Then we proceed as follows.

1. (E-step) We calculate the expectation of the log-likelihood $l(\lambda; X, Y)$ with respect to the conditional distribution of Y given X and under the current parameter $\zeta^{(n+1)}$,

$$Q(\eta|\zeta^{(n+1)}) = E_{Y|X,\zeta^{(n+1)}} l(\eta; X, Y).$$

2. (M-step) We find the maximizer of $Q(\eta|\zeta^{(n+1)})$ as a function of η,

$$\eta^{(n+1)} = \arg\max_{\eta} Q(\eta|\zeta^{(n+1)}).$$

The explicit form of optimizers in the EM algorithm is given below.
Using the result

$$\mathbb{E}_{q(\cdot)}\left[\log \lambda_{i,k}^{k'}\right] = \Psi\left(\gamma_{i,k}^{k'}\right) - \Psi\left(\sum_{s=1}^{K} \gamma_{i,s}^{k'}\right), \tag{60}$$

where $\Psi(\cdot)$ is the digamma function, the objective function $Q(\eta|\zeta)$ in the E-step is given by

$$
\begin{aligned}
Q(\eta|\zeta) &= \mathbb{E}_{q(\cdot)}[\log p(t, e, \mathbf{z}, \Lambda, \xi)|\zeta] \\
&= \sum_{i=1}^{m} \Bigg\{ \sum_{n=2}^{N_i} \Bigg[\sum_{k,k'=1}^{K} \tilde{\phi}_{i,n-1}^{(k',k)} \bigg(g_{k',k} - \frac{\tilde{a}_i}{\tilde{d}_i} \cdot e^{g_{k',k}} \big(t_{i,n} - t_{i,n-1} \big) + \log b_{k,e_{i,n}} \bigg) \Bigg] \\
&\quad + \quad (N_i + a - 2)\big[\Psi(\tilde{a}_i) - \log\big(\tilde{d}_i\big)\big] + a \log d - \log\Gamma(a) - d \cdot \frac{\tilde{a}_i}{\tilde{d}_i} \\
&\quad + \quad \sum_{n=2}^{N_i} \sum_{k,k'=1}^{K} \tilde{\phi}_{i,n-1}^{(k',k)} \Bigg(\Psi(\gamma_{i,k}^{k'}) - \Psi\Bigg(\sum_{s=1}^{K} \gamma_{i,s}^{k'} \Bigg) \Bigg) + \sum_{k=1}^{K} \phi_{i,1}^{(k)} \big(\log p_k^0 + \log b_{k,e_{i,1}} \big) \\
&\quad + \quad \sum_{k'=1}^{K} \sum_{k=1}^{K} \Bigg((r_k^{k'} - 1) \Bigg(\Psi(\gamma_{i,k}^{k'}) - \Psi\Bigg(\sum_{s=1}^{K} \gamma_{i,s}^{k'} \Bigg) \Bigg) \Bigg) \\
&\quad + \sum_{k'=1}^{K} \Bigg(\log\Gamma\Bigg(\sum_{k=1}^{K} r_k^{k'} \Bigg) - \sum_{k=1}^{K} \log\Gamma(r_k^{k'}) \Bigg) \Bigg\}.
\end{aligned}
\tag{61}
$$

In the M-step, we separate terms and maximize with respect to each parameter. The corresponding objective functions are

$$
Q(B) = \sum_{i=1}^{m} \sum_{n=1}^{N_i} \sum_{k=1}^{K} \sum_{v=1}^{V} \phi_{i,n}^{(k)} \mathrm{I}\{e_{i,n} = v\} \log b_{k,v},
$$

$$
Q(R) = \sum_{i=1}^{m} \Bigg\{ \sum_{k'=1}^{K} \sum_{k=1}^{K} \Bigg(r_k^{k'} \cdot \Bigg(\Psi(\gamma_{i,k}^{k'}) - \Psi\Bigg(\sum_{k=1}^{K} \gamma_{i,k}^{k'} \Bigg) \Bigg) \Bigg) \Bigg) + \sum_{k'=1}^{K} \Bigg(\log\Gamma\Bigg(\sum_{k=1}^{K} r_k^{k'} \Bigg) - \sum_{k=1}^{K} \log\Gamma)(r_k^{k'}) \Bigg) \Bigg\},
$$

$$
Q(G) = \sum_{i=1}^{m} \sum_{n=2}^{N_i} \sum_{k,k'=1}^{K} \tilde{\phi}_{i,n-1}^{(k',k)} \bigg(g_{k',k} - \frac{\tilde{a}_i}{\tilde{d}_i} \cdot e^{\lambda_{k',k}} \big(t_{i,n} - t_{i,n-1} \big) \bigg),
$$

$$
Q\big(\mathbf{p}^0\big) = \sum_{i=1}^{m} \sum_{k=1}^{K} \phi_{i,1}^{(k)} \log p_k^0,
$$

$$
Q(a) = \sum_{i=1}^{m} \big[\Psi(\tilde{a}_i) - \log(\tilde{d}_i) + \log d \big] \cdot a - m \cdot \log\Gamma(a),
$$

$$
Q(d) = m \cdot a \log d - d \cdot \sum_{i=1}^{m} \frac{\tilde{a}_i}{\tilde{d}_i}.
$$

The derivatives of $Q(B)$, Q(λ) and $Q(\mathbf{p}^0)$ are given by

$$
\frac{\partial Q(B)}{\partial b_{k,v}} = \frac{1}{b_{k,v}} \sum_{i=1}^{m} \sum_{n=1}^{N_i - 1} \phi_{i,n}^{(k)} \mathrm{I}\{e_{i,n} = v\} - \frac{1}{b_{k,V}} \sum_{i=1}^{m} \sum_{n=1}^{N_i} \phi_{i,n}^{(k)} \mathrm{I}\{e_{i,n} = V\},
$$

$$\frac{\partial Q(\lambda)}{\partial \lambda_{k',k}} = \sum_{i=1}^{m} \sum_{n=2}^{N_i} \tilde{\phi}_{i,n-1}^{(k',k)} - e^{\lambda_{k',k}} \sum_{i=1}^{m} \frac{\tilde{a}_i}{\tilde{d}_i} \sum_{n=2}^{N_i} \tilde{\phi}_{i,n-1}^{(k',k)} \left(t_{i,n} - t_{i,n-1}\right),$$

$$\frac{\partial Q(\mathbf{p}^0)}{\partial p_k^0} = \sum_{i=1}^{m} \phi_{i,1}^{(k)} \frac{1}{p_k^0} - \sum_{i=1}^{m} \phi_{i,1}^{(K)} \frac{1}{p_K^0},$$

$$\frac{\partial Q(a)}{\partial a} = \sum_{i=1}^{m} \left[\Psi(\tilde{a}_i) - \log\left(\tilde{d}_i\right)\right] + m \cdot \log d - m \cdot \Psi(a),$$

$$\frac{\partial Q(d)}{\partial d} = \frac{m \cdot a}{d} - \sum_{i=1}^{m} \frac{\tilde{a}_i}{\tilde{d}_i}.$$

We set the derivatives above to 0, and the corresponding optimizers have closed forms

$$b_{k,v} = \frac{\sum\limits_{i=1}^{m} \sum\limits_{n=1}^{N_i} \phi_{i,n}^{(k)} \mathrm{I}\{e_{i,n} = v\}}{\sum\limits_{i=1}^{m} \sum\limits_{n=1}^{N_i} \phi_{i,n}^{(k)}}, \tag{62}$$

$$g_{k',k} = \log \frac{\sum\limits_{i=1}^{m} \sum\limits_{n=2}^{N_i} \tilde{\phi}_{i,n-1}^{(k',k)}}{\sum\limits_{i=1}^{m} \frac{\tilde{a}_i}{\tilde{d}_i} \sum\limits_{n=2}^{N_i} \tilde{\phi}_{i,n-1}^{(k',k)} \left(t_{i,n} - t_{i,n-1}\right)}, \tag{63}$$

$$p_k^0 = \frac{\sum\limits_{i=1}^{m} \phi_{i,1}^{(k)}}{m}, \tag{64}$$

$$d = \frac{m \cdot a}{\sum\limits_{i=1}^{m} \tilde{a}_i / \tilde{d}_i}. \tag{65}$$

We update a by gradient descent. As for $Q(R)$, we calculate its first and second derivatives and use the Newton–Raphson algorithm to get the optimizers. The derivatives are

$$\frac{\partial Q(R)}{\partial r_s^k} = \sum_{i=1}^{m} \left(\Psi(\gamma_{i,s}^k) - \Psi\left(\sum_{l=1}^{K} \gamma_{i,l}^k \right) \right) + m \left(\Psi\left(\sum_{l=1}^{K} r_l^k \right) - \Psi(r_s^k) \right),$$

$$\frac{\partial^2 Q(R)}{\partial \alpha_s^k \partial \alpha_l^{k'}} = m \left(\mathrm{I}\{k = k'\} \cdot \Psi^{(1)} \left(\sum_{r=1}^{K} r_r^k \right) - \mathrm{I}\{k = k', s = l\} \cdot \Psi^{(1)}(r_s^k). \right)$$

We denote the gradient vectors and Hessian matrices as

$$g_\alpha(\mathbf{r}^k) = \left(\frac{\partial Q(R)}{\partial r_s^k}\right)_{K\times 1}, \quad k = 1\ldots K,$$

$$H_\alpha\left(\mathbf{r}^k\right) = \left(\frac{\partial^2 Q(R)}{\partial r_s^k \partial r_l^k}\right)_{K\times K}, \quad k = 1\ldots K.$$

In the (n+1)th iteration of the Newton–Raphson method, the estimates are updated as

$$\mathbf{r}_{(n+1)}^k = \mathbf{r}_{(n)}^k - H_\alpha(\mathbf{r}_{(n)}^k)^{-1} g_\alpha\left(\mathbf{r}_{(n)}^k\right). \tag{66}$$

We decompose the matrix $H_\alpha(\cdot)$ as

$$H_\alpha\left(\mathbf{r}^k\right) = m\left(D\left(\mathbf{r}^k\right) + c_k \cdot 1 \times 1^T\right),$$

where

$$D\left(\mathbf{r}^k\right) = \operatorname{diag}\left\{-\Psi^{(1)}\left(\alpha_1^k\right), \ldots, -\Psi^{(1)}\left(\alpha_K^k\right)\right\},$$

$$c_k = \Psi^{(1)}\left(\sum_{s=1}^{K} \alpha_s^k\right).$$

We can apply the matrix inversion lemma and get

$$m \cdot H_\alpha(\mathbf{r}^k)^{-1} = D(\mathbf{r}^k)^{-1} - \frac{D(\mathbf{r}^k)^{-1} 1 \times 1^T D(\mathbf{r}^k)^{-1}}{c_k^{-1} + \sum_{s=1}^{K} (d_s^k)^{-1}},$$

where d_s^k is the sth diagonal element of $D(\mathbf{r}^k)$. We let g_s^k denote the sth element of $g_\alpha(\mathbf{r}^k)$ and

$$\tilde{c}_k = \frac{\sum_{s=1}^{K} g_s^k / d_s^k}{c_k^{-1} + \sum_{s=1}^{K} (d_s^k)^{-1}},$$

so now

$$\left(H_\alpha(\mathbf{r}^k)^{-1} g_\alpha(\mathbf{r}^k)\right)_s = \frac{g_s^k - \tilde{c}_k}{m \cdot d_s^k}.$$

We then plug this into the equation (66) to get the parameter updates.

Appendix D:

Notation and assumptions

In this section we list the notation and assumptions that appear in the main text.

- Let $\eta = (B, G)$ for notational simplicity and let η^* be the true model parameters.

- Let $y_n = (e_n, t_n)$ and $\mathbf{Y} = (y_1, y_2, \ldots,)$. Let $\mathbf{Y}_i = (y_{i1}, \ldots, y_{iN_i})$ be an independent copy of $\mathbf{Y}$.
- Under bounded duration setting, it is supposed that $\tau_i \overset{i.i.d.}{\sim} f_\tau . f_\tau$ is some density function with bounded support in $\mathbb{R}^+$.
- We define $EL_{\tau,b}(\eta) = \mathbb{E}_{\eta^*} f(Y|\eta)$, where the expectation of Y is taken under true parameter $\eta^* = (B^*, G^*)$, and $f(Y|\eta) \equiv \max_q\{\mathbb{E}_q \log p(Y|\eta) f_\tau(\tau) - \mathbb{E}_q \log q\}$.
- Define $\check{\eta}(\tau)$ to be arg $\max_\eta EL_{\tau,b}(\eta)$, which represents the best approximate parameter under the proposed variational family.
- Define

$$A_1(\tau) = \mathbb{E}_{\eta^*}\left(\left.\frac{\partial f(Y|\eta)}{\partial \eta}\right|_{\check{\eta}(\tau)}\right)\left(\left.\frac{\partial f(Y|\eta)}{\partial \eta}\right|_{\check{\eta}(\tau)}\right)^T,$$

$$A_2(\tau) = \mathbb{E}_{\eta^*}\left.\frac{\partial^2 f(Y|\eta)}{\partial \eta^2}\right|_{\check{\eta}(\tau)}.$$

- In the large duration setting, it is supposed that each each individual has a true underlying personal transition probability Λ_i^* which defines an aperiodic and irreducible Markov chain and has a true underlying personal frailty ξ_i^*.
- Let $l_\tau(\eta, \mathbf{Y}) = \frac{1}{\tau}\log P(\mathbf{Y}|\eta)$ and $l_{i,\tau}(\eta, \mathbf{Y}_i) = \frac{1}{\tau}\log P(\mathbf{Y}_i|\eta)$.
- Let $g_n(\eta, \mathbf{Y}) = \log P(y_0|y_{-1}, \ldots, y_{-n})$ and $g(\eta, \mathbf{Y}) = \lim_{n\to\infty} g_n(\eta, \mathbf{Y}_i)$. Let $g_{i,n}(\eta, \mathbf{Y}_i) = \log P(y_{i,0}|y_{i,-1}, \ldots, y_{i,-n})$ and $g_i(\eta, \mathbf{Y}_i) = \lim_{k\to\infty} g_{i,n}(\eta, \mathbf{Y}_i)$, which are the sample versions of $g_n(\eta, \mathbf{Y})$ and $g(\eta, \mathbf{Y})$, respectively.
- Let $s_{\Lambda,\xi} = \lim_{\tau\to\infty}\left(N_{\Lambda,\xi}/\tau\right)$, representing the response speed. Let $s_i = \lim_{\tau\to\infty}(N_i/\tau_i)$, representing the individual version.
- Let $H_{\Lambda,\xi}(\eta, \xi) = E_{\eta^*_{\Lambda,\xi}} g(\eta, \mathbf{Y})$. Here, $\eta^*_{\Lambda,\xi} = (\Lambda, \xi, G^*, B^*)$ and the expectation of $\mathbf{Y}$ is taken under $\eta^*_{\Lambda,\xi}$. Further, we let $H_a(\eta) = \int s_{\Lambda,\xi} H_{\Lambda,\xi}(\eta) p(\Lambda) p(\xi) \mathrm{d}\Lambda \mathrm{d}\xi$.

Furthermore, we specify the following detailed assumptions.

A1. (Compactness) Suppose B and G lie on a compact parameter space. That is, $b_{k,e} \in [a', 1 - a']$ and $G_{k,k'} \in [a, A]$ for all k, k', e.

A2. The support of Λ's prior distribution is a compact set $\Theta_c \in \{\mathcal{S}_J\}^J$, where

$$\mathcal{S}_J = \left\{(\theta_1, \ldots, \theta_J) | \sum_j \theta_j = 1\right\}.$$

The support of ξ's prior distribution is a compact subset of $(0, +\infty)$.

A3. (Local identifiability) Both matrices $A_1(\tau)$ and $A_2(\tau)$ are of full rank.

A3$'$ (Local identifiability) $H_{\Lambda,\xi}(\eta)$ has three continuous-time derivatives with respect to η for all Λ and ξ. Let

$$Q_1 = \int \left(\frac{\partial s_{\Lambda,\xi} H_{\Lambda,\xi}(\eta)}{\partial \eta}\right)\left(\frac{\partial s_{\Lambda,\xi} H_{\Lambda,\xi}(\eta)}{\partial \eta}\right)^T p(\Lambda)p(\xi)\mathrm{d}\Lambda\mathrm{d}\xi,$$

$$Q_2 = \int \frac{\partial^2 s_{\Lambda,\xi} H_{\Lambda,\xi}(\eta)}{\partial \eta^2} p(\Lambda)p(\xi)\mathrm{d}\Lambda\mathrm{d}\xi,$$

evaluated at η^*. In fact, $Q_1 = Q_2$ We denote both of them by Q, assumed to be invertible.
A4$'$ (Exchangeability) We have

$$\begin{aligned}\lim_{\tau,m\to\infty} \frac{1}{m}\sum_i h_{i,\tau}(\eta,\mathbf{Y}_i) &= \lim_{\tau\to\infty}\lim_{m\to\infty}\frac{1}{m}\sum_i h_{i,\tau}(\eta,\mathbf{Y}_i)\\ &= \lim_{m\to\infty}\frac{1}{m}\sum_i \lim_{\tau\to\infty} h_{i,\tau}(\eta,\mathbf{Y}_i),\end{aligned}$$

where $h_{i,\tau}(\eta,\mathbf{Y}_i)$ could be $l_{i,\tau}(\eta,\mathbf{Y}_i)$, $\partial l_{i,\tau}(\eta,\mathbf{Y}_i)/\partial\eta$, $\partial^2 l_{i,\tau}(\eta,\mathbf{Y}_i)/\partial\eta^2$ or $\partial^3 l_{i,\tau}(\eta,\mathbf{Y}_i)/\partial\eta^3$.
A5$'$ $m \to \infty$ and $\tau_i = O(m^{r_0})$ for some $r_0 > 1$ for all i.

Algorithm 1:

Forward-backward variational EM algorithm

```
Input      : $t$, $e$.
Output     : Parameter estimates $\eta = \{B, G, \boldsymbol{p}^0, R, a, d\}, \zeta = \{\phi, \tilde{\phi}, \gamma, \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{d}}\}$.
Initialize: $p_{i,k}^{k'} (i = 1 : m, k = 1 : K, k' = 1 : K), \eta$.
while $Q(\eta|\zeta)$ has not converged do
    for $i \in \{1, \ldots, m\}$ do
        // Update forward and backward probabilities
        for $k \in \{1, \ldots, K\}, n \in \{1, \ldots, N_i\}$ do
            Update $f_{i,n}(k)$ using equation (47) and (48) ;
            Update $b_{i,N_i+1-n}(k)$ using equation (55) and (54) ;
        end
        // Update the posterior probabilites for latent topics
        for $k \in \{1, \ldots, K\}$, $n \in \{1, \ldots, N_i\}$ do
            Set $\phi_{i,n}^{(k)}$ by equation (34) ;
            for $l \in \{1, \ldots, K\}$ do
                Set $\tilde{\phi}_{i,n}^{(k',k)}$ by equation (35) ;
            end
        end
        // Update variational parameters $\boldsymbol{\gamma}_i^k$ and transition probabilities
        for $k \in \{1, \ldots, K\}$ do
            Set $\boldsymbol{\gamma}_i^k$ as in equation (27) ;
            for $k' \in \{1, \ldots, K\}$ do
                Set $p_{i,k}^{k'}$ by equation (28) ;
            end
        end
        Update $\tilde{a}_i, \tilde{d}_i$ by (29), (30) and update $\kappa_i$ ;
    end
    // We have updated all the parameters in $\zeta = \{\phi, \tilde{\phi}, \gamma, \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{d}}\}$
    // Then we apply the EM algorithm
    // E-step
    Get function $Q(\eta|\zeta)$ with updated $\zeta$ defined in equation (61) ;
    // M-step
    for $k \in \{1, \ldots, K\}$ do
        Set $p_k^0$ as (64) ;
        Optimize $Q$ function with respect to $\mathbf{r}^k$ ;
        for $v \in \{1, \ldots, V\}$ do
            Update $b_{k,v}$ by (62) ;
        end
        for $k' \in \{1, \ldots, K\}$ do
            Update $g_{k',k}$ by (63) ;
        end
    end
    Compute $Q(\eta|\zeta)$ with updated $\eta = \{B, G, \boldsymbol{p}^0, R, a, d\}$ and $\zeta = \{\phi, \tilde{\phi}, \gamma, \tilde{\boldsymbol{a}}, \tilde{\boldsymbol{d}}\}$ in
     equation (61) ;
end
```