Keystroke Logging

- Keystroke logging generates rich information about students' writing process that is typically not available in the final submitted products of writing tasks.
- We are interested in understanding students' writing process over the time of the writing task.

Logging Data

- The logging system captures keyboarding actions, e.g. deletion, insertion, and pause, along with their timestamps.
- The current position. Not considered in this study.
- Previous research used pre-specified rules to classify sequences of keyboarding actions into states such as text production, editing, long pause, etc. which are then modeled.
- Classification of pre-specified rules could be uncertain.
- We model the keyboarding actions directly and infer the latent states.

Observed Data

For a student, we observe J time durations.

- J may differ for different students, i.e. they spent different amount of time on the task.
- The choice of the length of the duration is subjective.
- We chose 5 seconds.

For the jth duration, there are K time units.

- The choice of K may depend on logging system capabilities or other constraints.

A keyboarding action is observed at the kth time unit, $X_{jk} = x_{jk}$, $x_{jk} \in \{\text{deletion, insertion, pause}\}$.

Latent States

A keyboarding action at the kth time unit, $X_{jk} = x_{jk}$, may be observed for different reasons, e.g.

- text production
- editing
- planning.

These states are latent and may be manifested in observed keyboarding actions.

- M latent states. M is a modeling choice.
- For the mth state, a multinomial distribution over the three possible observed keyboarding actions with a probability vector $\boldsymbol{\beta}_m$.
- For each X_{jk}, there is a latent state indicator $Z_{jk} = z_{jk}$, $z_{jk} \in \{1, 2, \ldots, M\}$.

Model 1: with augmented latent variables

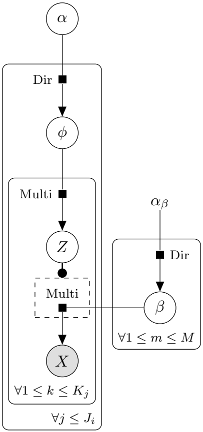


Figure: A mixed membership model with augmented discrete latent variables

$$\phi_j \sim \text{Dirichlet}(\boldsymbol{\alpha}),$$
$$Z_{jk} \sim \text{Multinomial}(\phi_j),$$
$$\boldsymbol{\beta}_m \sim \text{Dirichlet}(\boldsymbol{\alpha}_\beta),$$
$$X_{jk} \sim \text{Multinomial}(\boldsymbol{\beta}_{z_{jk}}).$$

- With a large number of Z_{jk}, MCMC could show slow mixing.
- We can sum Z_{jk} out of the model.
- Then we have $\boldsymbol{X}_j$ being a vector of counts over possible observed actions within the jth time duration.

Model 2: without the discrete latent variables

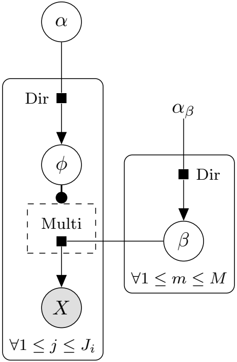


Figure: A mixed membership model with discrete latent variables marginalized out

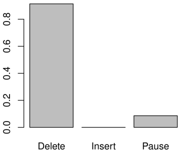
$$
\begin{aligned}
\boldsymbol{\phi}_j &\sim \text{Dirichlet}(\boldsymbol{\alpha}), \\
\boldsymbol{\beta}_m &\sim \text{Dirichlet}(\boldsymbol{\alpha}_{\boldsymbol{\beta}}), \\
\boldsymbol{X}_j &\sim \sum_k \phi_{jk} \text{Multinomial}(\boldsymbol{\beta}_k), \\
\boldsymbol{X}_j &\sim \text{Multinomial}(\boldsymbol{\beta}\boldsymbol{\phi}_j').
\end{aligned}
$$

- Technically, we are approximating the observed multinomial probability masses at each time duration with a mixture of multinomials.
- The mixture proportions could vary over different durations, thus providing a flexible and potentially better approximation.
- With the discrete latent variable summed out, it should lead to better mixing in MCMC.

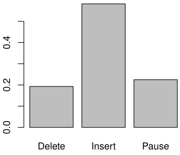
Empirical Example

- A high school student spent about 20 mins responding to a writing task.
- Timestamped keyboarding actions: Delete, Insert, and Pause.
- At 10 milliseconds (0.01s) level, Each time unit is associated with one of the three actions.
- Every 5 seconds constitute a time duration. So each duration has 500 observations except the last one.
- Fit the model with MCMC.

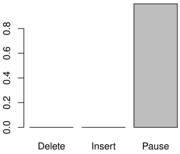
Results



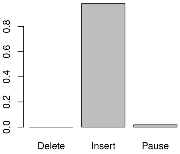
(a) Class 1



(b) Class 2



(c) Class 3

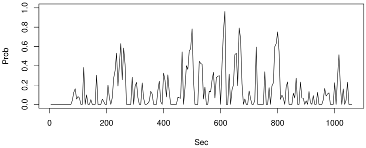


(d) Class 4

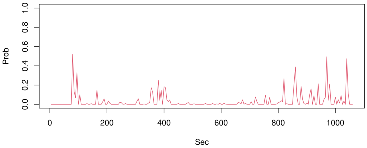
Figure: Distribution of actions by latent classes

Interpretation of Latent Classes

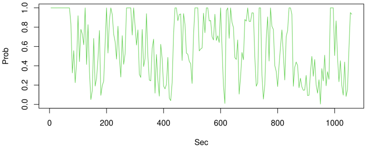
- Class 1, delete: deleting texts
- Class 2, constipation: Inserting(producing) texts slowly interspersed with deleting actions.
- Class 3, pause: no keyboarding actions.
- Class 4, text production: quickly inserting texts.



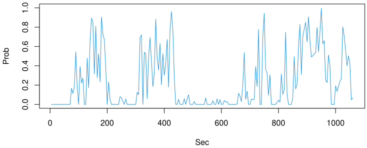
(a) Delete



(b) Constipation



(c) Pause



(d) Text Production

Figure: Mixture proportions over time

- Editing behaviors.
- Possible cognitive channel overload during spikes in constipation state.

Discussion

- Model identification issues. The likelihood of the model is invariant to permutation of labels of mixture components.
- Opportunity to integrate out mixing proportions with a Dirichlet prior.
- Generalizing the method for a group of students.
- Detailed feedback of writing process to students and teachers.
- Individualized learning opportunity.
- Real-time intervention.