# Mixed Membership Modeling of Student Strategies from Sequences of Actions

April Galyardt*

**Abstract**

Strategy choice strongly distinguishes novice and expert performance; however identifying strategies from data is an open problem. We propose a mixed membership-Markov chain (MM-MC) model to model how students use strategies. Markov processes can model the probabilistic sequence of actions that a student using a particular strategy will take, and a mixed membership framework will allow us to model students switching strategy from task to task. An earlier model, the simplicial mixture of Markov chains (SM-MC), was unsuccessful for data sets with N=1500, a fairly large sample size for educational data. SM-MC includes an exchangeability assumption equivalent to allowing students to switch strategy between every action. MM-MC simplifies the model by only allowing students to switch strategies between tasks. Since this assumes that short sequences of actions must have all come from the same Markov process, this makes it easier to estimate the transition matrices, and makes the model more tractable. We also include results from preliminary simulation studies.

**Key Words:** Mixed Membership, Latent Dirichlet Allocation, Markov Chain, Simplicial Mixture, Cognitive Strategies

## 1. Introduction

Currently there are no statistical or psychometric models that can capture student strategy usage. We seek to develop a statistical model that can be used to discover the strategies that students use from data, and eventually assess student expertise. The foundation for this model consists of three claims from cognitive science: 1) Strategy is a key feature that distinguishes expert and novice performance. 2) In most settings, people do not use a single strategy for every task, rather they switch strategy from task to task. 3) Sequences of actions contain information about strategies.

We use the example of adding numbers when both addends are less than 5 to illustrate these three claims. There are three basic addition strategies; 1) *retrieval* or memorization, 2) *count-on*, where to add 3+2, the child counts three-four-five, and 3) *count-all*, where to add 3+2, the child counts one-two-three-four-five. There are variations on these strategies such as counting on from the largest number or counting on from the first number, in addition to a guessing strategy, but for the purposes of this example will focus on these three strategies.

Simply reaching a correct answer does not distinguish experts from novices, since all three strategies will yield a correct answer. But we would generally consider the person who can quickly retrieve 3+2=5 from memory as more of an expert than a person who has to count. Indeed, 5-year olds are likely to need to start counting at one, and to have trouble starting at a different number. But 7-year olds are much more likely to use a mixture of retrieval and count-on [7]. The transition from

*University of Georgia, Athens, GA, 30602

count-all to count-on marks an important change in mathematical understanding. Thus strategy here is a critical difference between novice and expert.

This leads us to the second claim; Siegler [7] found that 99% of children used a mixture of strategies for adding 2 numbers less than ten. Campbell, et al.[1] found that 50% of adults also used a mixture of strategies for adding 2 numbers less than five. This phenomenon of switching from one strategy to another between tasks has been documented in a wide variety of domains, including arithmetic [1, 7], spelling [6], and mental rotation tasks [8].

If students who used one strategy always used that strategy, then there would be no significant modeling challenge. The strategy profiles would be latent classes, and we would be able to describe student behavior with a mixture model. However, since students switch strategies from problem to problem, we need a model that can capture this switching behavior. Mixed membership models have this capability [2, 3].

The final claim is that sequences of actions can reveal strategies. This proceeds directly from the definition of strategy. If a strategy is a plan of action, then observable behaviors taken to execute that plan should contain information allowing us to draw conclusions about the strategy.

Once we conceptualize a strategy as a likely sequence of actions, Markov processes are strong candidates for modeling this behavior. However, we cannot use a basic Markov chain or even a mixture of Markov processes, because students switch strategies. That is, on one problem, they may choose strategy Z which has a particular likely sequence of actions. On the next problem, they may choose another strategy W with a different likely sequence of actions.

The mixed membership-Markov chain (MM-MC) model addresses these issues. Each strategy is represented as a single Markov process, and the mixed membership structure allows for individuals using a mixture of strategies across different tasks. An existing model, the simplicial mixture of Markov chains (SM-MC) [5] has a similar structure. However, SM-MC has been shown to fail to recover parameters for moderately large data sets [4]. Briefly, SM-MC fails because it is too flexible. MM-MC makes a small change in the exchangeability assumptions which provides an estimable model. After formally introducing MM-MC in section 2, we discuss the relationship between the two models in section 2.1. Section 3 demonstrates MM-MC on simulated data.

## 2. Mixed Membership - Markov Chain Model

First, we assume that there is a single set of metacognitive strategies that is common across all students, indexed $k = 1, \ldots, K$. Second, we assume that each student, $i = 1, \ldots, N$, may use these strategies in different proportions. For example, some students may be more likely to guess when they don't know an answer, while other students may be more likely to ask for a hint. How much each student uses each strategy is parameterized by the vector $\theta_i = (\theta_{i1}, \ldots, \theta_{iK})$, so that $\theta_{ik}$ is the proportion of problems that student $i$ uses strategy $k$.

When a student begins a task $r = 1, \ldots, R$, they will choose (consciously or unconsciously) a strategy $Z_{ir} \in \{1, \ldots, K\}$ which will determine the likely sequence of actions.

$$Pr(Z_{ir} = k|\theta_i) = \theta_{ik} \tag{1}$$

or, equivalently,

$$Z_{ir}|\theta_i \sim Multinomial(\theta_i). \tag{2}$$

Note that the strategy $Z_{ir}$ depends only on the person $i$, not the item $r$. This is a simplification which seems unlikely to be true; different tasks may tend to elicit different strategies. Future work will explore the interaction between students and items in generating the strategy choice $Z$.

Once the student chooses the strategy $Z_{ir}$, they will take a sequence of actions $X_{ir} = (X_{ir1}, \ldots, X_{irt}, \ldots, X_{irT_{ir}})$ as they try to complete the task. Note that the length of the sequence, $T_{ir}$, differs from student to student and item to item.

Each strategy is defined by a discrete time Markov process. The state-space for the Markov chain is the set of observable student actions. Each Markov chain $k = 1, \ldots, K$ is parameterized by the initial probability vector $\pi_k$, and the transition probability matrix $P_k$. Thus, the probability of a student's sequence of actions $X_{ir}$ given their strategy choice $Z_{ir}$ is modeled as:

$$Pr(X_{ir} = x | Z_{ir} = k) \;=\; \pi_{k,x_1} \prod_{t=2}^{T_{ir}} P_{k x_{t-1} x_t} \quad . \tag{3}$$

Thus,

$$\begin{aligned} Pr(X_{ir} = x | \theta_i) &= \sum_{k=1}^{K} Pr(Z_{ir} = k | \theta_i) Pr(X_{ir} = x | Z_{ir} = k) && (4) \\ &= \sum_{k=1}^{K} \theta_{ik} \left[ \pi_{k,x_1} \prod_{t=2}^{T_{ir}} P_{k x_{t-1} x_t} \right], && (5) \end{aligned}$$

and finally, if we denote $X_i = (X_{i1}, \ldots, X_{ir}, \ldots, X_{iR})$, we have,

$$Pr(X_i = x | \theta_i) \;=\; \prod_{r=1}^{R} \left[ \sum_{k=1}^{K} \theta_{ik} \left[ \pi_{k,x_{r1}} \prod_{t=2}^{T_{ir}} P_{k x_{r(t-1)} x_{rt}} \right] \right]. \tag{6}$$

Note that these equations are written using a first-order Markov process, but they are easily extensible to higher order processes.

### 2.1 Comparison to Simplicial Mixtures of Markov Chains

Simplicial mixtures of Markov chains (SM-MC) [5] may be considered a special case of MM-MC. In SM-MC, $T_{ir} = 1$ for all $r$, so that individuals are modeled as having an opportunity to switch strategy profiles between each action. This assumption allows for a mathematical simplification:

$$Pr(X_{it} = s | \theta_i, X_{i,t-1} = s') = \sum_{k=1}^{K} \theta_{ik} P_{k,ss'} = P_{i,ss'} \tag{7}$$

In this special case, there is an individual transition matrix $P_i$ which is a convex combination of the profile transition matrices, $P_k$. Thus, SM-MC allows us to interpret individual strategies as a blend of the profiles [3]. Since this simplification is not possible in MM-MC, only the 'switching' interpretation is available: that is, students switch from using one strategy on one item to another strategy on the next item. However, this is exactly the cognitive behavior we are trying to capture.

Galyardt & Goldin [4] found that SM-MC could not recover parameters with $N = 1500$, $T_i = 200$, when there are only 3-5 profiles. The posterior distribution

of each of the transition matrices $P_k$ collapsed to a global mean transition matrix. The posterior distribution of $\theta$ collapsed to a point-mass with $\theta_i = \frac{1}{K}$ for all $i$.

We believe that this model is too flexible for the purpose of modeling student strategies. By assuming that students might switch strategies between each action, it is difficult to attribute any single transition to a particular profile transition matrix. MM-MC restricts the model by assuming that short sequences of actions all came from the same strategy profile.

## 3. Simulation

In this preliminary test of the model, we examined our ability to recover parameter estimates for data generated by the MM-MC model. To make the problem relatively easy for this preliminary test, we generated profile distributions that were fairly distinct from each other, and used a distribution of the membership parameters concentrated in the corners of the simplex. However, we also used a very small sample, to truly test the utility of this model in educational settings, where $N = 1000$ is rather large.

We generated 100 datasets using first-order Markov processes for the profiles, K=4 profiles, N=50 students, and J=15 tasks per student. The number of actions per task was generated according to Poisson($\lambda = 4$), so that we observed an average of 60 actions per student.

The number of total states in the Markov process, $S = 5$, is the number possible actions a student might take. The profile transition matrices were randomly generated: each row was drawn independently from a symmetric Dirichlet distribution, $P_{k,s\cdot} \sim \text{Dirichlet}(0.5)$, the initial state probabilities $\pi_k$ were generated from the same distribution. Membership parameters were generated according to $\theta_i \sim \text{Dirichlet}(0.25)$.

We estimated the model using MCMC. The prior distributions used for $\theta_i$, $P_{kss'}$ and $\pi_k$, were the generating distributions. For initialization of the MCMC draws, the initial state probability $\pi_k$ was randomly generated from the prior distribution. Then the initial profile assignments $Z_{ij}$ were drawn based solely on the first state in the sequence $X_{ij1}$ and the generated $\pi_k$. Other parameters were then updated in turn. In subsequent MCMC iterations, the $Z_{ij}$'s were updated using the full sequence, $X_{ij}$.

In all 100 simulations, we accurately recovered the transition matrices. Generating and estimated transition matrices from a randomly chosen simulation are shown in figure 2. To get a sense of how well we succeed in recovering transition probabilities across all simulations, we use posterior risk, that is, how close is the posterior distribution on average to the generating parameter. Let $P^{(b)}_{kss'}$ denote an MCMC draw for $P_{kss'}$. Then an estimate of posterior risk is:

$$\frac{1}{B}\sum_{b=1}^{B}\left(P^{(b)}_{kss'} - P_{kss'}\right)^2. \tag{8}$$

Figure 1 shows that we recovered the majority of transition probabilities with high accuracy, but that there are a few outliers with higher risk. We note that we are working with a relatively small sample size, 50 students with an average of 60 actions per student, means that we observe a total of around 3000 transitions in each simulation. With samples of this size, some states may be rarely observed in some of the profiles, leading to higher uncertainty in transition probatilities from those states. Thus, outliers in the distribution of risk are not surprising.

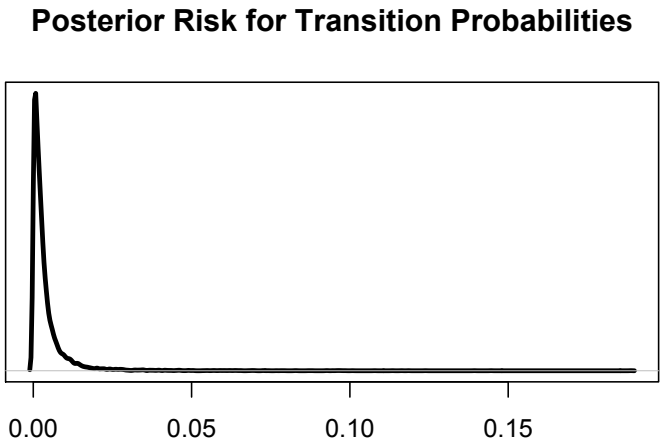


**Figure 1**: The distribution of posterior risk for $P_{kss'}$ across all simulations.

## 4. Discussion

These initial results indicate that MM-MC is a feasible model for discovering student strategies. We are able to accurately capture the differences in strategies with small data sets, 50 students and 15 tasks per student is a very modest amount of data that can feasibly be gathered by researchers in education.

The ability to discover student strategies automatically from data is useful for fundamental research in education and cognitive science. However, we have not yet demonstrated that MM-MC is useful as an *assessment* model. More work is necessary to investigate how many tasks per student are required to accurately estimate the membership parameters $\theta_i$. Future work must also address how the required size of $N$ and $J$ increase as the number of states $S$ increases.

The performance of MM-MC relative to SM-MC is remarkable. SM-MC was unable to recover any parameters with sample sizes of $N = 1500$, and 200 actions per individual, for a total of 300,000 observed actions [4]. With MM-MC we get accurate estimation with a data set that is 1% of that size, a mere 3,000 total observed actions. The only difference is a slight change in the exchangeability assumptions. MM-MC assumes that short sequences of observed actions came from the same profile. For the purposes of modeling student strategies, this is a very reasonable assumption.

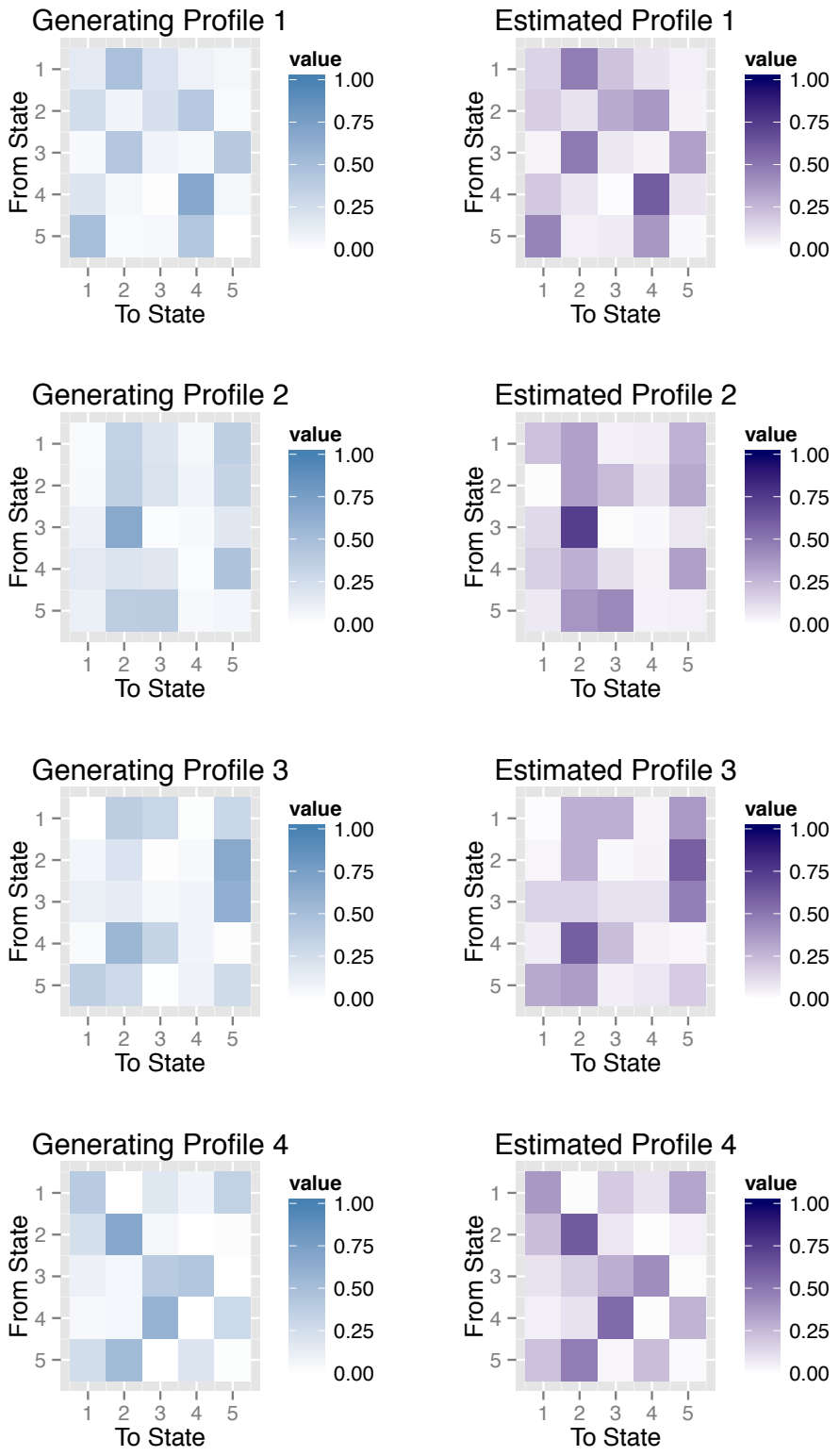


**Figure 2**: Heat maps of profile transition matrices for a randomly chosen simulation. The estimated transition matrices show posterior median probabilities.

## References

[1] Jamie I.D. Campbell and Shauna Austin. Effects of response time deadlines on adults' strategy choices for simple addition. *Memory & Cognition*, 30(6): 988–994, 2002.

[2] April Galyardt. *Mixed Membership Distributions with Applications to Modeling Multiple Strategy Usage*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA 15213, July 2012.

[3] April Galyardt. Interpreting mixed membership models: Implications of erosheva's representation theorem. In Edoardo M. Airoldi, David Blei, Elena Erosheva, and Stephen E. Fienberg, editors, *Handbook of Mixed Membership Models*. Chapman and Hall, 2014.

[4] April Galyardt and Ilya Goldin. Modeling student metacognitive strategies in a intelligent tutoring system. In *International Meeting of the Psychometric Society*, Arnhem, Netherlands, 2013.

[5] Mark Girolami and Ata Kaban. Sequential activity profiling: Latent dirichlet allocation of markov chains. *Data Mining and Knowledge Discovery*, 10:175–196, 2005.

[6] Sarah P. McGeown, Emma Medford, and Gerri Moxon. Individual differences in children's reading and spelling strategies and the skills supporting strategy use. *Learning and Individual Differences*, 28:75–81, 2013.

[7] Robert S. Siegler. The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General*, 116(3): 250–264, 1987.

[8] Irene Strasser, Ingrid Koller, Sabine Strauss, Mathis Csisinko, Hannes Kaufmann, and Judith Gluck. Use of strategy in a 3-dimensional spatial ability test. *Journal of Individual Differences*, 31(2):74–77, 2010.