Analyzing Transitions in Sequential Data with Marginal Models

Jeffrey Matayoshi
McGraw Hill ALEKS
jeffrey.matayoshi@mheducation.com

Shamya Karumbaiah
Carnegie Mellon University
shamya@cmu.edu

Various areas of educational research are interested in the transitions between different states—or events—in sequential data, with the goal of understanding the significance of these transitions; one notable example is affect dynamics, which aims to identify important transitions between affective states. Unfortunately, several works have uncovered issues with the metrics and procedures commonly used to analyze these transitions. As a recent study has revealed a statistical bias with several metrics used in sequential data analysis, we begin by looking in more detail at the specific mechanisms that are responsible for this bias. After giving a theoretical explanation for the issue, we present an alternative procedure that attempts to address the problem with the use of marginal models. A related issue is that the common practice of removing transitions to repeated states has been shown to have unintended side-effects, causing the importance of certain transitions to be overestimated—to account for this issue, we outline a method for extending the marginal model procedure to this type of analysis. Finally, in a recent study evaluating the problem of multiple comparisons and sequential data analysis, the Benjamini-Hochberg (BH) procedure, a commonly used approach to control for false discoveries, did not perform as expected. By applying a technique from the biostatistics and epidemiology literature, we show that performance of the BH procedure can be brought back to its expected level. In all of our analyses, we evaluate the proposed procedures by both running simulations and using actual student data. The results indicate that the marginal model procedure seemingly compensates for the bias observed in other transition metrics, thus resulting in more accurate estimates of the importance of transitions between states.

Keywords: sequential data, transition metrics, marginal models, affect dynamics

1. Introduction

As learning is a process that occurs over time, many areas of education and learning analytics research require the analysis of data that have a sequential or temporal ordering. Such analyses are important, as our understanding of the learning process can be greatly improved by leveraging the temporal features of these data (Knight et al., 2017). Additionally, properly analyzing the sequential properties of educational data has been shown to help improve the performance and accuracy of student models (Andrade et al., 2017; Mahzoon et al., 2018). When dealing with sequential data, researchers are often interested in the transitions that occur between different states—or events—in these sequences. One prominent example is in the area of affect dynamics, with the goal being to identify transitions between affective states that are highly significant (D'Mello and Graesser, 2012; Karumbaiah et al., 2018). Other works have applied

similar analyses to logs of student actions in learning systems, in an attempt to understand how students transition between different activities within these systems (Biswas et al., 2010; Bosch and D'Mello, 2017).

Unfortunately, multiple issues have recently been uncovered with the analysis of transitions in sequential data. For example, the work by Bosch and Paquette (2021) evaluated several metrics commonly used to analyze transitions within sequential data. In addition to looking at the probability estimates of different transitions, the study also evaluated techniques and transition metrics such as lag sequential analysis (Sackett, 1979) and the L statistic (D'Mello et al., 2007). Using numerical simulations, the analysis revealed a subtle statistical bias that occurs with these transition metrics, causing them to return unexpected and inflated values. This bias then creates extra difficulties when interpreting the values of the transition metrics, thus making it harder to measure the significance of transitions; additionally, the experiments in Bosch and Paquette (2021) showed that this issue is especially pronounced in short sequences of transitions.

A special case that has also turned out to be problematic is the handling of *self-transitions* in sequential data; these are simply transitions where the student remains in the same state for more than one step in a sequence. In many recent studies, researchers have removed self-transitions before analyzing the data with the L statistic (see the review in Karumbaiah et al. 2018 for further information). However, Karumbaiah et al. (2019) showed that excluding self-transitions has unintended consequences when used in combination with the L statistic, most likely giving misleading results. Thus, recent work in this area has focused on addressing this issue by either suggesting a modified interpretation of the L statistic values (Karumbaiah et al., 2019; Karumbaiah et al., 2021), or by using an altered version of the L statistic (Bosch and Paquette, 2021; Matayoshi and Karumbaiah, 2020).

Yet another issue occurs when dealing with the problem of multiple comparisons. Specifically, when evaluating transitions in sequential data, many pairs of transitions are typically analyzed with statistical tests and, as such, the probability of making a *discovery*—i.e., rejecting a null hypothesis—is higher than in an analysis involving a single statistical test. Thus, it follows that the probability of rejecting a true null hypothesis increases as well; such errors are variously called *false positives*, *false discoveries*, or *type I errors*. This is known in the statistics literature as the multiple comparisons problem. When analyzing sequential data, it's common to apply the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) to control the false discovery rate (FDR). One complication with using the BH procedure is that, in order for the theoretical guarantees on its performance to hold, the statistical tests must either be independent or satisfy certain dependency conditions (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). However, the difficulty is that it is not always straightforward to verify the conditions for applying the BH procedure; while some scenarios have been mathematically proven to satisfy these conditions, many common examples have not been. Thus, Matayoshi and Karumbaiah (2021a) ran numerical experiments evaluating the performance of the BH procedure when applied to the analysis of sequential data, and their results indicated the BH procedure doesn't always perform as expected in such situations, leading to a higher-than-expected FDR.

Motivated by all of the aforementioned issues, in this work we outline and evaluate a complete procedure for analyzing transitions in sequential data. To start, our analysis in Section 2, which was originally presented in Matayoshi and Karumbaiah (2021b), addresses the bias with transition metrics discovered by Bosch and Paquette (2021). After first replicating the Bosch and Paquette (2021) numerical experiments, we give a theoretical analysis that attempts to explain the underlying mechanisms causing the observed statistical bias. Based on this explanation, we

then outline a regression procedure that measures the significance of transitions using a marginal model approach. To evaluate the effectiveness of this procedure, we apply it to the simulated data generated in our numerical experiments.

Next, the analysis in Section 3, which is new and has not appeared in the literature previously, extends the marginal model procedure to the case when self-transitions are removed. We present a series of theoretical results that outline the proposed model and show that the at chance values—that is, the values under the assumption of complete independence—returned by the model are unbiased. We then evaluate the performance of the model in numerical experiments on simulated data.

To address the issues with multiple comparisons and applying the BH procedure to sequential data, in Section 4 we extend the work started in Matayoshi and Karumbaiah (2021a). In particular, we describe and evaluate adjustments to the marginal model approach that are taken from the biostatistics and epidemiology literature. Along with observing that these adjustments improve the performance of the BH procedure, we also find that, when self-transitions are excluded, the marginal model gives improved performance in comparison to L^*, a version of the L statistic specifically modified for the case when self-transitions are removed (Matayoshi and Karumbaiah, 2020). Finally, we consolidate all of these ideas in Section 5 and evaluate the complete procedure on real student data.

2. Transition Metrics and Statistical Bias

In this section we investigate a statistical bias with the analysis of state transitions that was first uncovered by Bosch and Paquette (2021). The work in this section originally appeared in Matayoshi and Karumbaiah (2021b).

2.1. Transition Metric Simulations

Consider the case when transitions between states happen purely at chance; that is, at all times in a sequence of states, the next state is sampled uniformly at random from all possible states. In such a case, we want our transition metric to return a baseline value that indicates the transitions are happening randomly and are not influenced by the starting state. This is the setting for the numerical experiments in Bosch and Paquette (2021), and we begin our current analysis with a replication of their work. To that end, consider two possible states, A and B. For each sequence length from 3 to 150, we generate 10,000 sequences of the given length by choosing between A and B at random; that is, we randomly choose between A and B with an equal probability of 0.5. For each of these sequences, we compute the values of $P(B \mid A)$ and $P(A \mid A)$; the former is the probability of transitioning to B, given that the starting state is A, while the latter is the probability of transitioning to A, given that the starting state is also A. Once we've computed these values for each sequence, we then compute the average for each conditional probability over the entire group of 10,000 sequences. Additionally, as another point of comparison, as done in Bosch and Paquette (2021) we also include the values from the L statistic, a popular transition metric used in the field of affective dynamics. The L statistic, which was originally introduced in D'Mello et al. (2007), is defined as follows.

Definition 1 (L statistic)**.** For states A and B, let $A \rightarrow B$ represent transitions that start in state

A and end in state B. We then have

$$L(A \to B) := \frac{P(B \mid A) - P(B)}{1 - P(B)}, \tag{2.1}$$

where $P(B)$ is the overall probability of B occurring as the next state and $P(B \mid A)$ is the conditional probability of transitioning to B, given that the starting state is A.

As with the conditional probabilities, we first compute the L values individually for each sequence, and we then find the averages of these values from the entire group of sequences. The values for both the conditional probabilities and the L statistic are shown in Figure 1, and it's worth noting that the results are consistent with those from Bosch and Paquette (2021). Regarding the conditional probabilities, while we expect these to be close to 0.5, as we are choosing between A and B equally at random, we can see that the computed values are heavily biased for the shortest sequences, with the bias then decreasing—but not completely disappearing—as the sequence length grows. In particular, the conditional probability values measuring transitions from $A \to A$ are biased in the negative direction, while the values for transitions of the form $A \to B$ show a bias in the positive direction. Turning next to the L statistic, we expect the values to be close to zero as, again, the states are being chosen uniformly at random. However, as with the conditional probabilities, we can see that there is a bias that is especially pronounced for the shortest sequences. For example, the maximum value of $L(A \to B)$ is just over 0.4 and the minimum value of $L(A \to A)$ is just under -0.5, and both of these values occur with the sequences of length 3. Note that, while the bias is fairly minimal once we reach sequence lengths of 40 or 50, obtaining this amount of data in a physical classroom can be challenging and impractical.[1]

2.2. Bias with Transition Metrics—Theoretical Analysis

Now that we have replicated the experiments from Bosch and Paquette (2021), we next offer a theoretical explanation for the biased values observed in these experiments.[2] The core of the issue can be traced to the conditional probability estimates of $P(B \mid A)$ and $P(A \mid A)$. Our claim is that the bias is an artifact of the averaging procedure used to estimate these values across the different sequences, and that this bias is then carried through to various transition metrics, such as L, that rely on these estimates. We illustrate the issue using a simple example. Consider the eight distinct sequences of length three consisting only of the states A or B, or both.

$$AAA \quad AAB \quad ABA \quad ABB \quad BAA \quad BAB \quad BBA \quad BBB$$

Now, as all transitions are equally likely in this set of sequences—i.e., all transitions occur with the same frequency—we would expect the computed estimates of $P(B \mid A)$ and $P(A \mid A)$

[1] As a simple example, consider a relatively small classroom containing 10 students. If we assume that the observation window is 20 seconds—which is fairly standard in affect dynamics research—it would take a single observer more than 2 hours to obtain 40 observations for each student, something that isn't possible when a class period is under an hour, as many are.

[2] After the work in Matayoshi and Karumbaiah (2021b) was completed, we became aware of research from Miller and Sanjurjo (2018) showing that a similar bias occurs when studying the "hot hand fallacy" of Gilovich et al. (1985)—that is, the question of whether or not a successful outcome makes it more likely for subsequent outcomes to be successful. While the original study did not find evidence for such a relationship (Gilovich et al., 1985), the work by Miller and Sanjurjo (2018) suggested that, after correcting for the statistical bias, there is evidence for the "hot hand" after all.

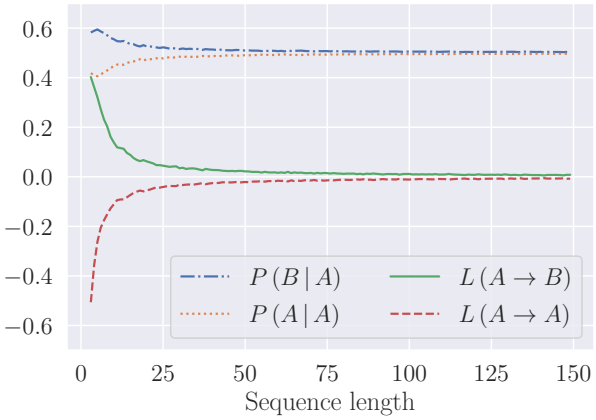


Figure 1: Plot of L values and conditional probabilities from our replication of the simulations in Bosch and Paquette (2021). The sequences are generated by choosing states A and B equally at random.

to each be 0.5. However, as shown by the values in the Unweighted column of Table 1 this is not the case. If we compute the probabilities individually for each sequence, and we then compute the averages over all the sequences, we obtain a value of 0.42 for $P(A \mid A)$ and a value of 0.58 for $P(B \mid A)$. In this example, the averaging procedure ignores the number of transitions that occur within each sequence, which then distorts the estimates. For example, the sequence AAA contains two transitions that start in A, while the sequence BAB contains only one; however, this discrepancy is ignored when computing the values in the Unweighted column of Table 1. Based on these results, this effect can be summarized, in some sense, by saying that high values of $P(A \mid B)$ occur more frequently than high values of $P(A \mid A)$ when the number of transitions within the sequences are ignored.

Next, consider what happens if, instead of averaging the conditional probabilities over the sequences, we compute the conditional probabilities by combining—or pooling—all of the data. That is, rather than grouping the transitions by sequence, we simply compute the rates of the transitions over the entire data set. Equivalently, we can also think of this as computing a weighted average of the conditional probabilities per sequence, where the weight is determined by the number of relevant transitions. For example, since sequences such as AAA and AAB contain two transitions that start in A, we assign these a weight of 2; on the other hand, sequences such as ABB and BAB only contain one transition from A, so these sequences are assigned a weight of 1. The results are shown in the Weighted column of Table 1, where we can see that the weighted conditional probabilities are both equal to 0.5, as desired.

2.3. Regression Procedure using Marginal Models

Based on the discussion in the previous section, the bias in the conditional probability estimates can be removed by using the extra information that is lost when averaging the values for each sequence. Thus, in what follows we describe a procedure that attempts to retain this information with the use of a logistic regression model. To begin, suppose we are interested in studying

Table 1: Computed weighted conditional probabilities.

	AAA	AAB	ABA	ABB	BAA	BAB	BBA	BBB	Mean	
									Unweighted	Weighted
$P(A \mid A)$	1	0.5	0	0	1	0	–	–	0.42	0.5
$P(B \mid A)$	0	0.5	1	1	0	1	–	–	0.58	0.5
Weight	2	2	1	1	1	1	0	0		

transitions of the form $A \to B$. Furthermore, assume that there are no restrictions on transitions between states.[3] To estimate the effect that starting in A has on transitions to B, we build a regression model in which the response variable is binary, with a value of one if the next state is equal to B, and a value of zero otherwise. Due to the binary form of the response variable, we use the logit as our link function. The sole predictor variable is another binary variable that is one if the previous state is equal to A, and zero otherwise. Under this formulation, a sequence of length n generates $n - 1$ data points. The variables of the model are summarized as follows.

- $y = y_{it}$: one if B is the next state for student i at time t; zero otherwise
- $x = x_{it}$: one if A is the previous state for student i at time t; zero otherwise

Letting σ represent the standard logistic function, the regression equation then has the form

$$P(y_{it} = 1 \mid x_{it}) = \sigma(\beta_0 + \beta_1 x_{it}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{it})}}. \tag{2.2}$$

In our analysis, we are interested in the coefficient β_1 from (2.2). The value of β_1 gives an indication of how much starting in state A—in comparison to not starting in state A—influences the likelihood of transitioning to state B. A large positive value for this coefficient would suggest that starting in A *increases* the likelihood of transitioning to B, in comparison to starting in a state that is not A. Analogously, a negative value for the coefficient might suggest that starting in A *decreases* the likelihood of transitioning to B, in comparison to not starting in A. Additionally, a benefit of this approach is that we can compare the probability estimates from the regression for the two different values of the predictor variable—zero or one—to get an additional measure of how large of an effect the predictor variable has.

One complication with the above procedure is that we have to deal with dependent—or correlated—data, as each sequence of transitions contains multiple measurements from the same student. Thus, when fitting the parameters for the logistic regression, we need to properly account for the dependence between these repeated measurements. We can accomplish this by using a multilevel model, where each individual student is considered a "group" or "cluster." Specifically, we use a marginal—or population averaged—model based on generalized estimating equations (GEE) (Heagerty and Zeger, 2000; Liang and Zeger, 1986). Marginal models are able to handle correlated data, and as such they are commonly used on data containing repeated measurements. We choose a marginal model because of our focus on estimating the average

[3] In comparison, some studies specifically ignore self-transitions—i.e., data points in which the same state appears consecutively—while in other situations certain transitions may be impossible. We return to this topic in Section 3.

response over the entire population, rather than estimating the effects on the individuals.[4] In order to account for the correlated data, we must specify the type of correlation structure for the data within each group. In our situation with repeated measurements, two common choices for the structure are an exchangeable correlation and a first-order autoregressive correlation. The exchangeable structure assumes that there is some common dependence between all the data in a group, while the autoregressive structure assumes that the dependence between the data in a group varies with time (Hardin and Hilbe, 2012; Heagerty and Zeger, 2000; Szmaragd et al., 2013). While it may occasionally be difficult to precisely determine the correct choice of correlation structure, it's worth noting that the parameter estimates are statistically consistent even if this structure is misspecified; in such a case, only the efficiency of these estimates is compromised (Hardin and Hilbe, 2012; Liang and Zeger, 1986).

Since the estimating equations used in GEE models are not necessarily likelihood based, we are unable to use the standard Akaike Information Criterion (AIC) (Akaike, 1974) to compare different models. Instead, we can compare the fits of different models using the Quasi-AIC (QIC) score (Pan, 2001); among other things, using the QIC score can help us determine the best choice of correlation structure (Hardin and Hilbe, 2012; Pan, 2001). Then, to analyze the effect of our predictor variable, we can evaluate β_1 using standard techniques such as the Wald test for statistical significance (Hardin and Hilbe, 2012). Another advantage of this approach is that it directly compares the cases when (a) the starting state is A and (b) the starting state is not A. In comparison, the L statistic compares the cases when the starting state is A to the overall behavior, regardless of the starting state. The drawback to the latter approach is that if A is very common and its occurrence dominates the sequence of states, it's possible that the values of $P(B)$ and $P(B \mid A)$ will be very close simply because A is almost always the starting state.

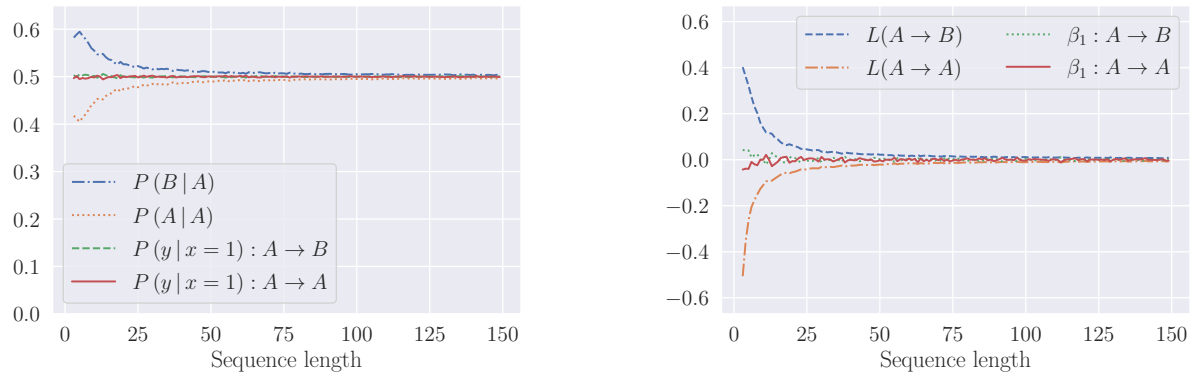
2.4. Experiments on Simulated Data

In this section we apply the marginal model approach to simulated data using the `statsmodels` (Seabold and Perktold, 2010) Python library.[5] We begin by applying the model to the data from our replication of the work in Bosch and Paquette (2021); recall that, for $n = 3, 4, \ldots, 150$, we generate 10,000 different sequences, each of length n, where each state in each sequence is chosen uniformly at random from A or B. The first set of results for the transitions $A \to A$ and $A \to B$ are shown in Figure 2a. There, we plot the unweighted conditional probability values, computed directly from the raw data and averaged over each set of 10,000 trials, along with the estimated probabilities from the marginal models; in the latter case—i.e., the dashed green line and solid red line—these estimates correspond to the model predictions when $x = 1$. We can see that, for both transition pairs $A \to A$ and $A \to B$, the estimates from the marginal models are all closely centered around 0.5; this is in sharp contrast to the computed conditional probability values, which exhibit the previously discussed bias.

Next, in Figure 2b we compare the L values with the values of β_1, the coefficient of our single predictor variable. As shown previously, the L values for $A \to B$ have a positive bias, with a maximum value of just over 0.4, while the L values for $A \to A$ have a negative bias, with a minimum value of just below -0.5. However, in all cases the β_1 values are closely

[4] If the focus is on the individuals, one possible approach is to estimate the subject-specific parameters by using a mixed-effects model with a random intercept for each student.

[5] A Python module for running all of our numerical experiments is available at `https://github.com/jmatayoshi/consolidated-transition-analysis`.



(a) (b)

Figure 2: Plots comparing (a) the unweighted conditional probability values and estimates from the marginal models, and (b) the L values and β_1 coefficients from the marginal models. The sequences are generated by choosing states A and B equally at random.

centered around zero, as is preferred. We should also mention that, for this analysis, we use the exchangeable correlation structure for the marginal models. As the states are chosen with equal probability from either A or B, there is actually no underlying dependence in the data; thus, it is instructive that, even with the incorrect correlation structure, the resulting parameter estimates are accurate.

To investigate the situation when the transition states occur with different frequencies—or base rates—we run one additional set of simulations. For these simulations, we assume there are four possible states: A, B, C, and D. To generate our sequences, we sample randomly according to the following distribution: A is chosen with probability 0.6, B is chosen with probability 0.2, and C and D are each chosen with probability 0.1. Then, for $n = 3, 4, \ldots, 150$, we generate 10,000 different sequences, each of length n, according to this probability distribution on the states. The results are shown in Figure 3, where we plot the computed conditional probabilities, along with the estimated probabilities from the marginal models. As before, we can see that the raw conditional probabilities are biased for the shorter sequences. In comparison, the estimates from the marginal models are centered closely around the true values.

3. Removing Self-Transitions

3.1. Modifying the Marginal Model Procedure

In this next section we extend the marginal model procedure to a particular situation that occurs in sequential data analysis. Specifically, we focus on the case when researchers want to remove the influence of repeated states. To do this, many researchers in the affect dynamics community remove *self-transitions*—i.e., transitions where the same state is repeated for more than one step—before analyzing the data (see Karumbaiah et al. 2018 for a review of recent works employing this technique). While this procedure appears logical at first glance, it has unintended consequences when analyzing the resulting sequences. To start, note that if states

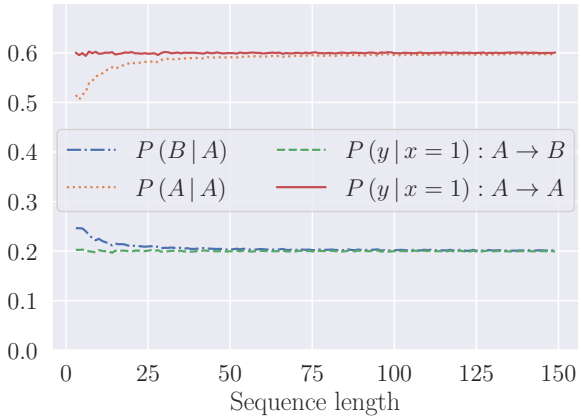


Figure 3: Plot of unweighted conditional probability values and the corresponding estimates from the marginal models. The sequences are generated by choosing A with probability 0.6, B with probability 0.2, and C and D each with probability 0.1.

A and B appear independently of each other then $P(B \mid A) = P(B)$; applying a measure such as the L statistic should then result in a value of zero, at least in theory (in practice, the statistical bias uncovered by Bosch and Paquette 2021 demonstrates this is not necessarily the case). However, as shown by Karumbaiah et al. (2019), removing self-transitions violates the assumption of independence between the appearances of A and B, as the next state can now only take on values other than A. As such, when self-transitions are excluded, under the assumption of independence the values of $P(B \mid A)$ and $P(B)$ are not necessarily equal, and we can no longer assume that zero represents the value of the L statistic in the independent case. To compensate for these issues, a modified version of the L statistic, named L^*, was introduced in Matayoshi and Karumbaiah (2020).

Definition 2. Let A and B be two states, and let

$$T_{\cdot,\overline{A}} = \{\text{transitions where the next state is not } A\}. \tag{3.1}$$

Then, we define

$$L^*(A \to B) := \frac{P\left(B \,\middle|\, A, T_{\cdot,\overline{A}}\right) - P\left(B \,\middle|\, T_{\cdot,\overline{A}}\right)}{1 - P\left(B \,\middle|\, A, T_{\cdot,\overline{A}}\right)}, \tag{3.2}$$

where $P(B \mid A, T_{\cdot,\overline{A}})$ is the probability of a transition to B in $T_{\cdot,\overline{A}}$, given that the starting state is A, while $P\left(B \,\middle|\, T_{\cdot,\overline{A}}\right)$ is the overall probability of a transition to B in $T_{\cdot,\overline{A}}$. The base rate of the state B, given by $P\left(B \,\middle|\, T_{\cdot,\overline{A}}\right)$ in (3.2), can be computed either individually for each sequence, or averaged over the entire set of sequences.

As discussed in Matayoshi and Karumbaiah (2020), the intuition behind the use of the set $T_{\cdot,\overline{A}}$ can be described as follows. Assume we have a transition that begins in affective state A. In order to reduce the influence of repeated transitions, suppose a researcher decides to exclude *all* self-transitions. Now, consider the comparison of the probabilities $P(B \mid A)$ and

$P(B)$. When self-transitions are excluded, $P(B \mid A)$ is computed with state A removed as a possible affective state to transition to—in contrast, $P(B)$ is computed under the scenario that all states are possible. The result is that, in most cases, this would serve to inflate the difference $P(B \mid A) - P(B)$ as, all else equal, the probability of a transition to B is higher when there are fewer possible states. Another important observation is that, with self-transitions removed, any consecutive states must be different, which means $P(B)$ cannot be much larger than 0.5; as before, this could again inflate the difference $P(B \mid A) - P(B)$

While Matayoshi and Karumbaiah (2020) provided both theoretical and empirical evidence validating the use of L^*—with the empirical evidence being based on both simulated and real data—there is as yet no modification to the L statistic that satisfactorily adjusts for the statistical bias discussed in the previous section. Thus, in the interest of having a uniform procedure that is applicable regardless of whether or not self-transitions are removed, we next outline how the above discussion for L^* can be used to extend the marginal model procedure to the case when self-transitions are removed.

To begin, observe that the formula for L^* (3.2) is equivalent to applying the formula for the L statistic (2.1) to the transitions in $T_{\cdot,\overline{A}}$. Thus, based on this intuition, the marginal model procedure can also be extended to the handle the removal of self-transitions by applying it to only the transitions in $T_{\cdot,\overline{A}}$. In this case, from (2.2) we then have

$$P(y_{it} = 1 \mid x_{it} = 1) = P(B \mid A, T_{\cdot,\overline{A}}).$$

That is, $P(y_{it} = 1 \mid x_{it} = 1)$ is the estimated probability of a transition to B in $T_{\cdot,\overline{A}}$, given that the starting state is A. Similarly, we have

$$P(y_{it} = 1 \mid x_{it} = 0) = P(B \mid \overline{A}, T_{\cdot,\overline{A}}).$$

That is, $P(y_{it} = 1 \mid x_{it} = 0)$ is the estimated probability of a transition to B in $T_{\cdot,\overline{A}}$, given that the starting state is not A. Note that these are normalized values of $P(B \mid A)$ and $P(B \mid \overline{A})$, in the sense that the influence of A (possibly) being the next state is removed. By doing this, our goal is to have the at chance value be centered at zero. To make this rigorous, we need to borrow the concept of *conditional independence*.

Definition 3. Let A and B be two affective states, and let $T_{\cdot,\overline{A}}$ be defined as in (3.1). Suppose that B_{next} represents the occurrence of a transition that ends in B, while A_{prev} represents the occurrence of a transition that starts in A. Then, we say that the events B_{next} and A_{prev} are conditionally independent given $T_{\cdot,\overline{A}}$ if

$$P(B_{next} \cap A_{prev} \mid T_{\cdot,\overline{A}}) = P(B_{next} \mid T_{\cdot,\overline{A}}) \cdot P(A_{prev} \mid T_{\cdot,\overline{A}}). \tag{3.3}$$

Note that the above definition of conditional independence is very similar to the standard definition of independence, with the only difference being that each probability is conditioned on $T_{\cdot,\overline{A}}$. Thus, if we restrict ourselves to looking only at transitions in $T_{\cdot,\overline{A}}$, the definition of conditional independence simplifies to the standard definition of independence, giving us a straightforward way to check if (3.3) holds.

The motivation for using conditional independence is the following. Suppose we have a transition of the form $A \to B$. As observed by Karumbaiah et al. (2019), when self-transitions are excluded there is no longer independence between the events A_{prev} and B_{next}, as the next state can now only take on values other than A; in other words, the set of possible values for

the next state is explicitly dependent on the value of the previous state. As a direct consequence of this dependence between the events A_{prev} and B_{next}, the equality of $P(B \mid A)$, $P(B \mid \overline{A})$, and $P(B)$ is no longer guaranteed, and thus we cannot assume that $L = 0$ or $\beta_1 = 0$ represent the values at chance.

Note that removing transitions to A has an effect on our concept of conditional independence—once transitions to A are removed, the formula for conditional independence (3.3) then simplifies to the formula for regular independence. Thus, if regular independence holds for the modified sequence with transitions to A removed, it follows that the probabilities $P(B \mid A)$, $P(B \mid \overline{A})$, and $P(B)$ should all be equal; that is, when transitions to A are excluded the concept of at chance is now captured by the conditional independence assumption. In order to formalize this result, we need the following lemma.

Lemma 1. Let A and B be two affective states, and let $T_{\cdot,\overline{A}}$ be defined as in (3.1). Assume that the events B_{next} and A_{prev} are conditionally independent given $T_{\cdot,\overline{A}}$. We then have

$$P(B_{next} \cap \overline{A}_{prev} \mid T_{\cdot,\overline{A}}) = P(B_{next} \mid T_{\cdot,\overline{A}}) \cdot P(\overline{A}_{prev} \mid T_{\cdot,\overline{A}}). \tag{3.4}$$

That is, B_{next} and $\overline{A}_{prev}$ are also conditionally independent given $T_{\cdot,\overline{A}}$.

Proof. We start by observing that

$$B_{next} \cap T_{\cdot,\overline{A}} = \left(B_{next} \cap \overline{A}_{prev} \cap T_{\cdot,\overline{A}}\right) \cup \left(B_{next} \cap A_{prev} \cap T_{\cdot,\overline{A}}\right),$$

and

$$\left(B_{next} \cap \overline{A}_{prev} \cap T_{\cdot,\overline{A}}\right) \cap \left(B_{next} \cap A_{prev} \cap T_{\cdot,\overline{A}}\right) = \varnothing.$$

Thus, it follows that

$$P\left(B_{next} \cap T_{\cdot,\overline{A}}\right) = P\left(B_{next} \cap \overline{A}_{prev} \cap T_{\cdot,\overline{A}}\right) + P\left(B_{next} \cap A_{prev} \cap T_{\cdot,\overline{A}}\right). \tag{3.5}$$

Next, rearranging the terms in (3.5) and applying the definition of conditional probability, we have

$$\begin{aligned}
P\left(B_{next} \cap \overline{A}_{prev} \,\middle|\, T_{\cdot,\overline{A}}\right) &= \frac{P\left(B_{next} \cap \overline{A}_{prev} \cap T_{\cdot,\overline{A}}\right)}{P(T_{\cdot,\overline{A}})} \\
&= \frac{P\left(B_{next} \cap T_{\cdot,\overline{A}}\right) - P\left(B_{next} \cap A_{prev} \cap T_{\cdot,\overline{A}}\right)}{P(T_{\cdot,\overline{A}})} \qquad \text{(using (3.5))} \\
&= \frac{P\left(B_{next} \cap T_{\cdot,\overline{A}}\right)}{P(T_{\cdot,\overline{A}})} - \frac{P\left(B_{next} \cap A_{prev} \cap T_{\cdot,\overline{A}}\right)}{P(T_{\cdot,\overline{A}})} \\
&= P\left(B_{next} \,\middle|\, T_{\cdot,\overline{A}}\right) - P\left(B_{next} \cap A_{prev} \,\middle|\, T_{\cdot,\overline{A}}\right).
\end{aligned}$$

Applying (3.3), it follows that

$$\begin{aligned}
&= P\left(B_{next} \,\middle|\, T_{\cdot,\overline{A}}\right) - P\left(B_{next} \,\middle|\, T_{\cdot,\overline{A}}\right) \cdot P\left(A_{prev} \,\middle|\, T_{\cdot,\overline{A}}\right) \qquad \text{(using (3.3))} \\
&= P\left(B_{next} \,\middle|\, T_{\cdot,\overline{A}}\right) \cdot \left(1 - P\left(A_{prev} \,\middle|\, T_{\cdot,\overline{A}}\right)\right) \\
&= P\left(B_{next} \,\middle|\, T_{\cdot,\overline{A}}\right) \cdot P\left(\overline{A}_{prev} \,\middle|\, T_{\cdot,\overline{A}}\right),
\end{aligned}$$

as claimed. □

We can now prove our main result.

Theorem 1. Let A and B be affective states, and suppose that the conditional independence requirement given in (3.3) holds. Assume also that $P\left(\overline{A}_{prev} \cap T_{\cdot,\overline{A}}\right) > 0$. Then,

$$P\left(B_{next} \,\middle|\, A_{prev}, T_{\cdot,\overline{A}}\right) = P\left(B_{next} \,\middle|\, \overline{A}_{prev}, T_{\cdot,\overline{A}}\right).$$

Proof. Starting with an application of the definition of conditional probability, we have

$$\begin{aligned}
P\left(B_{next} \,\middle|\, A_{prev}, T_{\cdot,\overline{A}}\right) &= \frac{P\left(B_{next} \cap A_{prev} \cap T_{\cdot,\overline{A}}\right)}{P\left(A_{prev} \cap T_{\cdot,\overline{A}}\right)} \\
&= \frac{P\left(B_{next} \cap A_{prev} \cap T_{\cdot,\overline{A}}\right)}{P\left(A_{prev} \cap T_{\cdot,\overline{A}}\right)} \cdot \frac{P\left(T_{\cdot,\overline{A}}\right)}{P\left(T_{\cdot,\overline{A}}\right)} \\
&= \frac{P\left(B_{next} \cap A_{prev} \,\middle|\, T_{\cdot,\overline{A}}\right)}{P\left(A_{prev} \,\middle|\, T_{\cdot,\overline{A}}\right)} \\
&\quad \text{(using the definition of conditional probability)} \\
&= \frac{P\left(B_{next} \,\middle|\, T_{\cdot,\overline{A}}\right) \cdot P\left(A_{prev} \,\middle|\, T_{\cdot,\overline{A}}\right)}{P\left(A_{prev} \,\middle|\, T_{\cdot,\overline{A}}\right)} \qquad \text{(using (3.3))} \\
&= P\left(B_{next} \,\middle|\, T_{\cdot,\overline{A}}\right).
\end{aligned}$$

Next, solving (3.4) for $P\left(B_{next} \,|\, T_{\cdot,\overline{A}}\right)$, it follows that

$$\begin{aligned}
&= \frac{P\left(B_{next} \cap \overline{A}_{prev} \,\middle|\, T_{\cdot,\overline{A}}\right)}{P\left(\overline{A}_{prev} \,\middle|\, T_{\cdot,\overline{A}}\right)} \\
&= \frac{P\left(B_{next} \cap \overline{A}_{prev} \cap T_{\cdot,\overline{A}}\right)}{P\left(T_{\cdot,\overline{A}}\right)} \cdot \frac{P\left(T_{\cdot,\overline{A}}\right)}{P\left(\overline{A}_{prev} \cap T_{\cdot,\overline{A}}\right)} \\
&\quad \text{(using the definition of conditional probability)} \\
&= \frac{P\left(B_{next} \cap \overline{A}_{prev} \cap T_{\cdot,\overline{A}}\right)}{P\left(\overline{A}_{prev} \cap T_{\cdot,\overline{A}}\right)} \\
&= P\left(B_{next} \,\middle|\, \overline{A}_{prev}, T_{\cdot,\overline{A}}\right),
\end{aligned}$$

as claimed.

□

3.2. Numerical Experiments

In this section we evaluate the marginal model procedure, adapted to remove the influence of self-transitions, on simulated data. Our first set of experiments is similar to the procedure applied in Section 2.4; as done there, for $n = 3, 4, \ldots, 150$ we generate 10,000 different sequences, each of length n. In this case, however, rather than limiting our sequences to only the states A and B, we instead select states uniformly at random from A, B, and C. Once this is done, we then restrict each sequence to the transitions in $T_{\cdot,\overline{A}}$ and apply the marginal model procedure. The resulting estimates of $P(B \,|\, A)$ are shown by the solid (blue) line in Figure 4, while Figure 5a

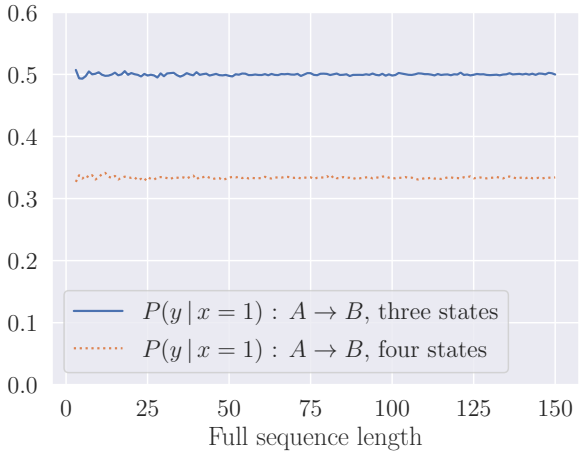


Figure 4: Results from using either three or four states chosen uniformly at random. The probability estimates are all centered around 0.5 or 0.33, as desired.

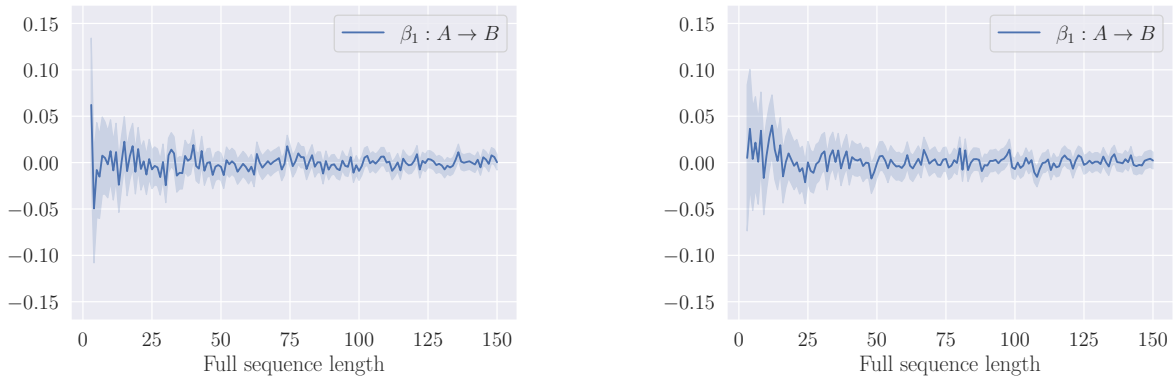
has the estimates for β_1. Then, for our second set of experiments we use a similar procedure, with the difference being that we select the states uniformly at random from A, B, C, and D. The corresponding estimates for $P(B \mid A)$ are shown by the dotted (orange) line in Figure 4, with Figure 5b then containing the estimates for β_1.

For the simulations with three total states (i.e., A, B, and C), the estimates of $P(B \mid A)$ in Figure 4 are all closely centered around 0.5, as desired. That is, given that we remove any transitions to A in order to estimate $P(B \mid A)$, we would expect B and C to appear roughly equally. Then, for the simulations with four total states (i.e., A, B, C, and D), the estimates of $P(B \mid A)$ in Figure 4 are centered around 0.33. This is the expected value, as we are removing transitions to A and we thus expect B, C, and D to each appear roughly equally. Next, the β_1 estimates in Figure 5a and Figure 5b, while somewhat noisy for the smaller sequence lengths, overall seem to be centered around zero. Additionally, the 95% confidence intervals for the point estimates of β_1 are represented by the shaded regions in the two plots. Each figure has 148 confidence intervals, and in both cases 141 of these confidence intervals contain 0, which works out to an estimated coverage probability of about 0.953. Thus, as we expect the coverage probability of a 95% confidence interval to be about 0.95, overall this suggests that the point estimates and confidence intervals for β_1 are reasonably accurate in both sets of simulations.

4. Controlling for Multiple Comparisons

Consider a statistical analysis that tests several different null hypotheses, either on related data sets or on a single data set. When following such a procedure, the probability of making a *discovery*—i.e., rejecting a null hypothesis—is higher than in an analysis involving a single null hypothesis. As such, it also follows that the probability of rejecting a true null hypothesis increases as well; such errors are variously called *false positives*, *false discoveries*, or *type I errors*. This is known in the statistics literature as the multiple comparisons problem.

In this section we evaluate the performance of the marginal model procedure in regards to false discoveries and the multiple comparisons problem. While doing so, we extend the work



(a) Three states

(b) Four states

Figure 5: Estimates of β_1 values from simulations using either three or four states. Shaded regions show the 95% confidence intervals.

from Matayoshi and Karumbaiah (2021a) by also evaluating the rate of false negatives, thereby giving a more complete analysis of the procedure's performance. Additionally, as Matayoshi and Karumbaiah (2021a) showed that in some cases the marginal model procedure has an inflated rate of false positives, we apply and evaluate an adjustment taken from the biostatistics literature that is intended to help correct for this issue (Li and Redden, 2015; Mancl and DeRouen, 2001). We use simulation studies for all of these evaluations, an approach that is commonly applied to investigate the performance of multiple comparison procedures (Benjamini, 2010; Benjamini and Hochberg, 1995; Farcomeni, 2006; Kim and van de Wiel, 2008; Reiner-Benaim, 2007; Reiner-Benaim et al., 2007; Williams et al., 1999; Yekutieli, 2008).

4.1. False Discovery Rate

The main focus of our current analysis is the false discovery rate (FDR). The FDR was introduced by Benjamini and Hochberg (1995), and it has since found widespread use in many scientific fields including education research (Williams et al., 1999), genetics (Reiner-Benaim, 2007; Storey and Tibshirani, 2003), and medical studies (Benjamini and Yekutieli, 2001). If we let $\mathbf{V}$ be the number of false discoveries and $\mathbf{S}$ be the number of true discoveries, as done in Benjamini and Hochberg (1995) we can define the quantity $\mathbf{Q}$ as

$$\mathbf{Q} = \begin{cases} \frac{\mathbf{V}}{\mathbf{V}+\mathbf{S}}, & \text{if } \mathbf{V}+\mathbf{S} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

Then, the FDR is equal to $\mathbf{E}[\mathbf{Q}]$, the expected proportion of false discoveries among all the discoveries made.

The family-wise error rate (FWER), which is defined as the probability of making at least one false discovery when performing a set of hypothesis tests, is another measure commonly associated with the problem of multiple comparisons. Although the Bonferroni correction is probably the most famous procedure used to control the FWER, there exist many other alternatives. However, while such procedures can be useful in situations in which a false discovery

must be avoided at all costs, such as clinical trials of new medical treatments (Goeman and Solari, 2014), the downside to these methods is a loss of statistical power, resulting in an increased likelihood of missing true discoveries. While procedures for controlling the FWER are concerned with the occurrence of *any* false discoveries, FDR controlling procedures are slightly more permissive, as they allow a certain proportion of the discoveries to be false. Thus, the advantage of FDR controlling procedures is that they typically have greater statistical power and, as such, a better chance of correctly identifying true discoveries; the resulting trade-off is that false discoveries are more likely. However, this trade-off can be beneficial when a large number of hypothesis tests are being conducted,[6] or when the research is of a slightly more exploratory nature.

In addition to introducing the FDR to the scientific literature, the authors in Benjamini and Hochberg (1995) also outlined what is now known as the Benjamini-Hochberg (BH) procedure. As shown there, the BH procedure is mathematically proven to control the FDR at a given level when the statistical tests—or, equivalently, the test statistics—are independent. However, in many practical applications the statistical tests may have some underlying dependence between them. With these situations in mind, further important work on controlling the FDR appeared in Benjamini and Yekutieli (2001), where the authors proved that, in addition to the independent case, the BH procedure is valid under certain dependency conditions between the statistical tests. Among other scenarios, it was shown that the BH procedure properly controls the FDR with multivariate normal test statistics having nonnegative correlations. Additionally, the authors in Benjamini and Yekutieli (2001) introduced a new method—now known as the Benjamini-Yekutieli (BY) procedure—for situations in which the BH procedure is not valid, and they proved that this new procedure controls the FDR regardless of the dependence between the tests.

In the remainder of this section we discuss the application of the BH and BY procedures. Consider a statistical analysis that involves the testing of m null hypotheses. Of these null hypotheses, $m_0 \leq m$ are true null hypotheses—these correspond to the hypotheses that we expect a statistical test to classify as not being significant—while the remaining $m - m_0$ hypotheses are the false null hypotheses. Note that, in practice, m_0 is an unknown value. Let $P_1, \ldots, P_m$ be the *p*-values for the m statistical tests, with these values being listed in ascending order; the corresponding null hypotheses are then represented by $H_1, \ldots, H_m$. The relationships between these various terms can be summarized as follows.

	Not significant	Significant	Total
True null	**U**	**V**	m_0
False null	**T**	**S**	$m - m_0$

(4.2)

- m = total number of hypotheses being tested
- m_0 = number of true null hypotheses
- **V** = number of false positives (i.e., false discoveries or type I errors)
- **S** = number of true positives
- **T** = number of false negatives (i.e., type II errors)
- **U** = number of true negatives

[6] As a relatively extreme example, statistical analyses in genetics research can involve thousands of hypothesis tests, and in such cases FWER controlling procedures can be overly restrictive (Benjamini, 2010).

Let q represent our chosen threshold—or, level—for controlling the FDR, and define the value $\text{FDR}_{\max} = \frac{m_0}{m}q$. If the statistical tests are independent, or if they satisfy certain dependency conditions, it was shown by Benjamini and Yekutieli (2001) that the FDR resulting from an application of the BH procedure is at most $\text{FDR}_{\max}$. Such an application works as follows. Assuming once again that the *p*-values are in ascending order, we find the largest integer k such that $P_k \leq \frac{k}{m}q$. Then, we simply reject all the null hypotheses H_i for which $i \leq k$.

Next, as the BY procedure controls the FDR under arbitrary dependence assumptions, it is necessarily more conservative when rejecting a null hypothesis. This takes the form of a lower threshold for the upper bound used to determine the "significance" of the *p*-values. Specifically, we find the largest integer k such that $P_k \leq \frac{k}{m \cdot c(m)}q$, where $c(m) = \sum_{i=1}^{m} \frac{1}{i}$. Using this procedure, it was shown in Benjamini and Yekutieli (2001) that the resulting FDR is bounded above by $\text{FDR}_{\max} = \frac{m_0}{m}q$, regardless of the type of dependence between the statistical tests.

To see how these procedures work, we next look at an example. Suppose we run 10 separate statistical tests ($m = 10$) that return the following *p*-values.

$$0.002,\ 0.008,\ 0.011,\ 0.013,\ 0.023,$$
$$0.028,\ 0.092,\ 0.214,\ 0.647,\ 0.853$$

Next, we compare these *p*-values to the formulas used for the BH and BY thresholds, using a value of $q = 0.1$; for added context, we also include the results for the Bonferroni correction. For each method, the thresholds that correspond to statistically significant *p*-values are in bold.

k	P_k	BH $\frac{k}{m}q$	BY $\frac{k}{m\sum_{i=1}^{m}\frac{1}{i}}q$	Bonferroni $\frac{1}{m}q$
1	0.002	**0.01**	**0.003**	**0.01**
2	0.008	**0.02**	**0.007**	**0.01**
3	0.011	**0.03**	**0.010**	0.01
4	0.013	**0.04**	**0.014**	0.01
5	0.023	**0.05**	0.017	0.01
6	0.028	**0.06**	0.020	0.01
7	0.092	0.07	0.024	0.01
8	0.214	0.08	0.027	0.01
9	0.647	0.09	0.031	0.01
10	0.853	0.1	0.034	0.01

For the BH procedure, we can see that $k = 6$ is the largest value for which $P_k \leq \frac{k}{m}q$, as we have $0.028 < 0.06$. Thus, the BH procedure, using a value of 0.1, would reject the null hypothesis for the statistical tests corresponding to the lowest six *p*-values. Next, for the BY procedure we see that $k = 4$ is the largest value for which P_k is less than the corresponding threshold; in this case, we have $0.013 < 0.014$. It's worth noting that, in this example, even though both P_2 and P_3 are *not* below the corresponding thresholds, the BY procedure still classifies them as being statistically significant. This is a feature of FDR controlling procedures that, in many cases, allows them to be more permissive than procedures for controlling the FWER.

4.2. METHODS

In this section we outline the general procedure we follow for our multiple comparison simulation studies. Since evaluating multiple comparison procedures requires knowledge of whether a null hypothesis is true, and as this isn't typically known with real data, simulations are commonly used for such analyses. In all of our experiments, we begin by generating simulated data according to a given probability distribution. While the specifics of this procedure vary slightly for our different experiments, the common thread is that this must be done in a way as to have control over whether or not each null hypothesis is true. Specifically, we have a single parameter that controls whether or not a subset of the states are related; thus, when this parameter is non-zero the null hypothesis that the states are independent is false.

Another important detail is that, as we are focusing on one particular scenario, we can generate simulated sequences of states specific to this scenario. By simulating the underlying data, we are attempting to evaluate the BH procedures in conditions that are as realistic as possible. In comparison, other studies that are more general in nature may simulate the distribution of the test statistics, rather than the underlying data, when evaluating multiple comparison procedures.

After generating the data for a simulation run, we perform our statistical tests and compute the corresponding p-values. Once this is done, we then apply the BH procedure for various threshold values q; specifically, we use 0.05, 0.1, and 0.15 in all our evaluations. While a value of 0.05 is commonly used, it's been argued that this threshold may be too low for some applications—for example, the sections on multiple comparisons and the FDR in James et al. (2021) and McDonald (2014) have useful discussions of this issue. As such, we evaluate a range of values in our simulations. Based on the statistical significance results from our applications of the BH procedure, we can compute $\mathbf{Q}$, the proportion of false discoveries among all the discoveries made, using (4.1). To obtain our estimate of the FDR, we then compute the average of $\mathbf{Q}$ over a total of 10,000 simulation runs. For the various values of q, we compare these FDR estimates to the values of FDR_{max} as defined in Section 4.1.

At this point, it's worth mentioning that the value of $\mathbf{Q}$—and, hence, the estimated FDR value—can be very different from the false positive rate.[7] Using the notation in (4.2), the false positive rate can be written as $\frac{\mathbf{V}}{\mathbf{V+U}}$. In comparison, $\mathbf{Q}$ is computed with the formula $\frac{\mathbf{V}}{\mathbf{V+S}}$, which has a different denominator. Thus, while the FDR is the expected proportion of false discoveries among all the rejected null hypotheses, the false positive rate is the (expected) proportion of false discoveries among all the true null hypotheses. Consider the following example. Assume we are testing 20 total hypotheses, all of which are true null hypotheses ($m_0 = m = 20$). Furthermore, assume that one false positive is recorded. Then, the false positive rate for this set of tests would be equal to $\frac{1}{1+19} = 0.05$. However, applying (4.1) gives a value of $\mathbf{Q} = \frac{1}{1+0} = 1$. This discrepancy is something to keep in mind as we analyze the results from our simulation studies in subsequent sections.

4.3. EXPERIMENTAL SETUP

Our numerical experiments for sequential data evaluate the BH procedure on simulated sequences of states. Each of these sequences could represent, for example, a student's affective states while working in a learning system. To generate a sequence of states, we start with the

[7] That is, while "false discovery" and "false positive" are used interchangeably, the terms "false discovery rate" and "false positive rate" have different definitions.

Table 2: Mean probability distribution used to generate the simulated sequences of states. Each entry represents the probability of making a transition to the next state (column), given the previous state (row). To simulate the differing behaviors of individual students, these starting rates are randomly adjusted each time a new sequence is generated—importantly, however, the distribution of the average rates matches what is shown in the table below.

prev \ next	A	B	C	D	E
A	0.2	$0.2+\gamma$	0.2	$0.2-\gamma$	0.2
B	0.2	0.2	0.2	0.2	0.2
C	0.2	$0.2-\gamma$	0.2	$0.2+\gamma$	0.2
D	0.2	0.2	0.2	0.2	0.2
E	0.2	0.2	0.2	0.2	0.2

probability distribution given in Table 2; each entry in this table gives the probability of sampling the next state (column) based on the value of the previous state (row). Note that while this is the distribution used for all of the simulations in Matayoshi and Karumbaiah (2021a), a possible criticism of this methodology is that using the exact same distribution to generate each sequence of states is not entirely realistic, as students likely have some variation in how often they transition to the different states. Thus, we use the following approach to simulate this variation at the student level. When generating each sequence of states we randomly sample a value α from the following list: 0.04, 0.08, 0.12. After obtaining α, we randomly select two states; for the first state we add α to each of the values in the corresponding column of Table 2, and then for the second state we subtract α from each of the values in the column corresponding to that state.

To help make the above explanation clearer, we next go through the process of generating an example sequence using a value of $\gamma = 0.05$. We begin by randomly selecting a value of α; suppose that $\alpha = 0.12$ is obtained. Next, we randomly select two states; for this, suppose that B and E are chosen, in that order. Thus, according to the procedure outlined in the previous paragraph, we first add $\alpha = 0.12$ to each of the values in the B column of Table 2; since $\gamma = 0.05$, the resulting values are 0.37, 0.32, 0.27, 0.32, and 0.32. Next, we subtract $\alpha = 0.12$ from each of the values in the E column of Table 2; as these values all start at 0.2, we end up with each value being shifted down to 0.08. The resulting probability distribution is shown in Table 3. These values can be interpreted as follows. Suppose that C is the previous state. In this case, A has a probability of 0.2 of being the next state, B has a probability of 0.32 of being the next state, and so on. Finally, in Table 4 we show the results from applying the marginal model procedure to a simulated sequence of 200 states, with each state being generated according to the probability distribution in Table 3.

For our simulations, we use two different values of γ: 0, which results in all 25 hypotheses being true null hypotheses; and 0.05, which results in 21 true null hypotheses, out of the 25. For each value of γ, we generate n sequences consisting of 20 states each. To generate these sequences, the first state in each sequence is sampled randomly from the five choices. Next, α and two states are randomly chosen, the distribution in Table 2 is updated based on these values, and then all subsequent states are sampled according to this new probability distribution. For each set of n sequences we evaluate our statistical tests (described in Sections 4.5 and 4.6) and

Table 3: Example probability distribution used to generate one sequence of simulated states, using $\gamma = 0.05$. In this example, all the base rates in column B have been shifted up by $\alpha = 0.12$, while all the base rates in column E have been shifted down by $\alpha = 0.12$.

prev \ next	A	B	C	D	E
A	0.2	0.37	0.2	0.15	0.08
B	0.2	0.32	0.2	0.2	0.08
C	0.2	0.27	0.2	0.25	0.08
D	0.2	0.32	0.2	0.2	0.08
E	0.2	0.32	0.2	0.2	0.08

Table 4: Marginal model coefficient p-values from one simulated sequence consisting of 200 states, with the states generated according to the probability distribution in Table 3. The bold p-values correspond to the four false null hypotheses that occur with a value of $\gamma = 0.05$—note that, as expected, these are by far the smallest p-values.

prev \ next	A	B	C	D	E
A	0.273	**0.000**	0.529	**0.000**	0.184
B	0.477	0.214	0.593	0.772	0.080
C	0.689	**0.012**	0.966	**0.000**	0.431
D	0.476	0.264	0.173	0.247	0.953
E	0.450	0.991	0.786	0.526	0.934

then compute the resulting value for **Q**; this constitutes one simulation run. We then perform 10,000 simulation runs for each value of n in order to obtain an estimate of the true FDR. For this analysis, we use the following values of n: 25, 50, 100, and 200.

As before, let m denote the total number of statistical tests, with $m_0 \leq m$ representing the number of true null hypotheses. Using the BH procedure with a value of $\gamma = 0$, we have $m_0 = m$; as such, we would expect the FDR to be less than $\text{FDR}_{\text{max}} = \frac{25}{25}q = q$ if the BH conditions are satisfied. Then, for all values of $\gamma > 0$ we would expect the FDR to be less than $\text{FDR}_{\text{max}} = \frac{21}{25}q$, assuming the BH conditions are satisfied, as $m_0 = 21$ of the tests are true null hypotheses.

4.4. Inflated Rate of False Positives

The simulation results in Matayoshi and Karumbaiah (2021a) suggested that using the BH procedure did not control the FDR at the expected rate. While it is an open question if the theoretical conditions for applying the BH procedure are satisfied, it was at least shown by Matayoshi and Karumbaiah (2021a) that the statistical tests are not independent of each other. However, setting aside the BH procedure for the moment, previous evaluations of marginal models have demonstrated that the rate of false positives can depend on the number of clusters—or, groups—appearing in the analysis, and some authors have suggested that at least 40 are needed for adequate performance and well-calibrated p-values (Mancl and DeRouen, 2001; Li and Redden,

2015). While the simulations in Matayoshi and Karumbaiah (2021a) used at least 50 clusters, our next analysis suggests that there are issues with the resulting p-values from these simulations.

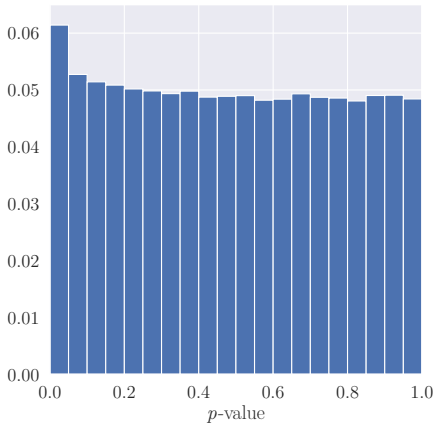
In Figure 6a we plot the p-values from a set of 10,000 simulation runs, with 50 sequences being generated in each run—as each sequence of states is considered a separate cluster, this means we have 50 clusters in each simulation run. For these simulations, the standard errors are computed using the "robust" setting, which is the default option of the `statsmodels` GEE class—these standard errors are derived using the so-called *sandwich estimator* of Liang and Zeger (1986). After obtaining the standard errors, these are divided into the regression coefficients, with the p-values then being computed using a standard normal distribution—note that this is the procedure used in the simulation study of Matayoshi and Karumbaiah (2021a). For this particular set of simulations we use a value of $\gamma = 0$, which means that only true null hypotheses are being tested—thus, for these simulations we want the p-values to be uniformly distributed from 0 to 1, as this ensures an accurate rate of false positives. For example, if we were to use a significance level of 0.05, we'd expect to incorrectly reject a true null hypothesis about 5% of the time. However, as shown in Figure 6a the p-values are not uniformly distributed, as there is an inflated amount of small p-values. Thus, such a bias would very likely explain the higher-than-expected FDR rates observed in Matayoshi and Karumbaiah (2021a).

In an attempt to correct for this bias, we borrow a technique from the biostatistics and epidemiology literature. The first step in applying this technique is to use the "bias reduced" option in the `statsmodels` GEE implementation—using this option returns standard errors that are computed with the adjustment outlined by Mancl and DeRouen (2001). Second, following the recommendation of Mancl and DeRouen (2001), the p-values are computed using a t-distribution with degrees of freedom equal to $N - k - 1$, where N is the number of clusters and k is the number of independent variables; in our case, we have $k = 1$. These adjusted p-values, which we refer to as MD-corrected p-values, have been shown to address the bias that appears with small numbers of clusters, and they have also been shown to work well when the sizes of the clusters vary (Mancl and DeRouen, 2001; Snijders and Bosker, 2012). As seen in Figure 6b, the resulting p-values are distributed much more evenly from 0 to 1 when compared to the p-values in Figure 6a. Thus, for our subsequent simulations concerning the marginal model procedure and multiple comparisons, we analyze the results using MD-corrected p-values.

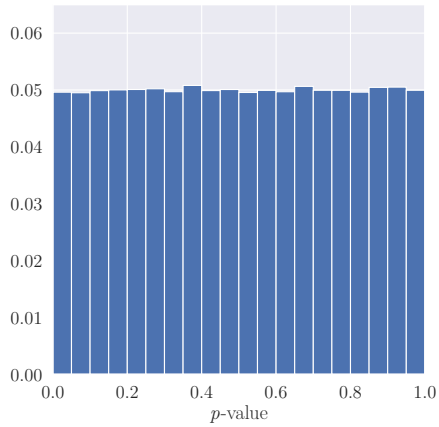
While MD-corrected p-values are available in many software implementations, we also investigated a simpler correction proposed by MacKinnon and White (1985). This correction simply scales the robust standard error values by the factor $\frac{N}{N-k-1}$; then, as with MD-corrected p-values, a t-distribution with $N - k - 1$ degrees of freedom is used to compute the p-values. The overall results were similar to those from using MD-corrected p-values. Thus, while MD-corrected p-values are generally recommended over this simpler correction (Angrist and Pischke, 2008; Mancl and DeRouen, 2001), in the absence of an available software implementation of the MD-correction, we believe that the simpler adjustment from MacKinnon and White (1985) is a reasonable alternative.

4.5. Results

In our next set of experiments we run the previously described numerical simulations and evaluate the performance of the BH procedure. Our first set of results are shown in Figure 7, where we use a value of $\gamma = 0$; that is, these simulations have no dependencies between the different



(a) Robust standard errors



(b) Standard errors with Mancl-DeRouen correction

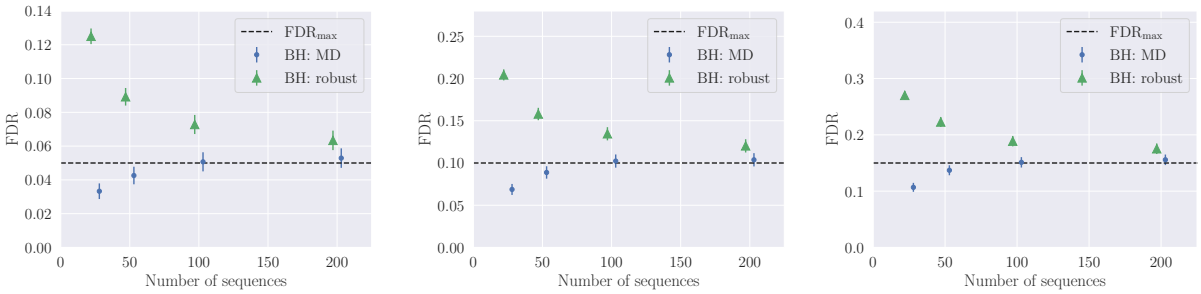
Figure 6: Relative frequency histograms comparing the p-values using (a) robust standard errors and a normal distribution and (b) standard errors with the Mancl-DeRouen correction and a t-distribution. The results in (a) show an overabundance of p-values less than 0.2, with an especially high number less than 0.05. In comparison, the values in (b) are distributed much more evenly from 0 to 1.

states, which means there are no false null hypotheses. In these results, we show the FDR estimates using the p-values computed with robust standard errors and a normal distribution (green triangles), along with the p-values using MD-corrected standard errors and a t-distribution (blue circles). In both cases the BH procedure is applied using various threshold values q. Overall, we can see that using MD-corrected p-values leads to much better control of the FDR, especially when the number of sequences is small. For example, using a q value of 0.1, the FDR estimates using the robust standard errors are as high as 0.2; contrast this with the results from using the MD-corrected p-values, where the highest FDR estimates just barely exceed the theoretical maximum of $\text{FDR}_{\max} = 0.1$.

In Figure 8 we show the results from our simulations using a value of $\gamma = 0.05$, which adds a dependence between 4 of the 25 pairs of states—this results in 4 false null hypotheses out of the 25 total hypotheses. As with the completely independent case, the MD-corrected p-values give much better control of the FDR, with the difference again being most pronounced when the number of sequences is small. Thus, combining the two sets of experiments in this section, the overall results suggest that the BH procedure performs well when used in combination with marginal models and MD-corrected p-values, thus addressing one of the issues reported by Matayoshi and Karumbaiah (2021a).

4.6. Excluding Self-Transitions

Our next analysis focuses on the situation when a researcher would like to remove the influence of self-transitions. To do this, we begin with the simulated sequences from our experiments in Section 4.5. Then, we take these sequences and apply our chosen model—for this analysis we compare the L^* statistic with the marginal model procedure outlined in Section 3.1. To test



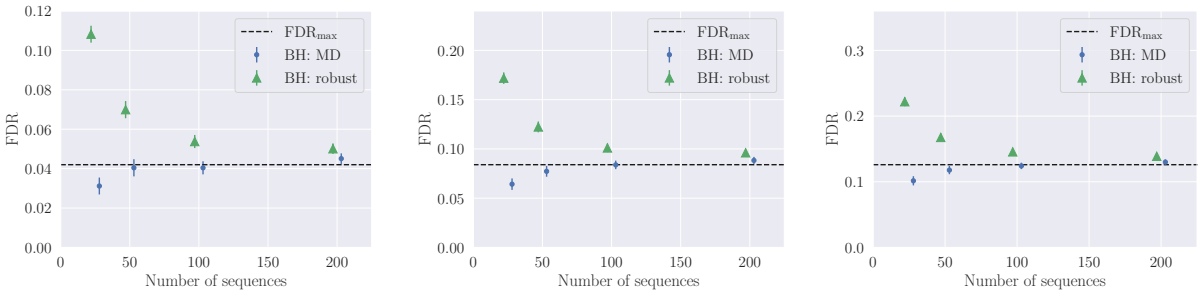
(a) $q = 0.05$ (b) $q = 0.1$ (c) $q = 0.15$

Figure 7: Comparison of the estimated FDR using a value of $\gamma = 0$; the BH procedure; and either robust standard errors with a normal distribution, or MD-corrected standard errors with a t-distribution. Vertical lines represent the 99% confidence interval for each estimated FDR value. To avoid overlapping points, the plotted x-values for the two methods—robust standard errors vs. MD-corrected standard errors—have been shifted by -3 and $+3$, respectively.

for statistical significance when using L^*, we follow the procedure outlined in Matayoshi and Karumbaiah (2020) and use a two-tailed t-test on the L^* values, and we then apply the BH procedure to the resulting p-values. The results for $\gamma = 0$ are shown in Figure 9. Note there are several examples where the estimated FDR values using L^* and the BH procedure are clearly above the FDR_{max} line. Similar to the results from our previous experiments that used the marginal model procedure and robust standard errors, the worst cases occur with the smallest number of sequences. In comparison, using the marginal model with MD-corrected p-values and the BH procedure gives good control of the FDR, with all the estimated FDR values falling below the theoretical FDR_{max} line.

We next evaluate the procedures when some dependence occurs between the states by using a value of $\gamma = 0.05$. Additionally, to give a more nuanced comparison between the L^* and marginal model procedures, we also evaluate the resulting true positive rates (TPR). To compute the TPR, we use the following approach. First, note that a value of $\gamma = 0.05$ results in 4 false null hypotheses out of the 20 total null hypotheses. Then, for each simulation run we compute the proportion of these four false null hypotheses that are classified as statistically significant after the BH procedure is applied. We then estimate the true value of the TPR by computing the average of these proportions over all of our 10,000 simulation runs.

The results evaluating both the FDR and TPR are shown in Figure 10. The top row of the figure shows the estimates for the FDR, where we can see that, as before, the FDR values from using L^* are typically higher than expected; in contrast, we can see that using the marginal model with MD-corrected p-values gives much stricter control of the FDR. Additionally, the TPR results in the second row of Figure 10 are illuminating. Given that the use of L^* leads to higher FDR values in all our evaluated cases, it is surprising that the marginal model TPR values are either comparable to the TPR values from L^*—such as with the smaller numbers of sequences—or substantially better than the TPR values for L^*. As an example of the latter case, with an x-value of 100 sequences and using $q = 0.1$, the estimated TPR value for the marginal model is about 0.63, much higher than the estimated value of 0.42 using L^*. Thus, the results



(a) $q = 0.05$ (b) $q = 0.1$ (c) $q = 0.15$

Figure 8: Comparison of the estimated FDR using a value of $\gamma = 0.05$; the BH procedure; and either robust standard errors with a normal distribution, or MD-corrected standard errors with a t-distribution. Vertical lines represent the 99% confidence interval for each estimated FDR value. To avoid overlapping points, the plotted x-values for the two methods—robust standard errors vs. MD-corrected standard errors—have been shifted by -3 and $+3$, respectively.

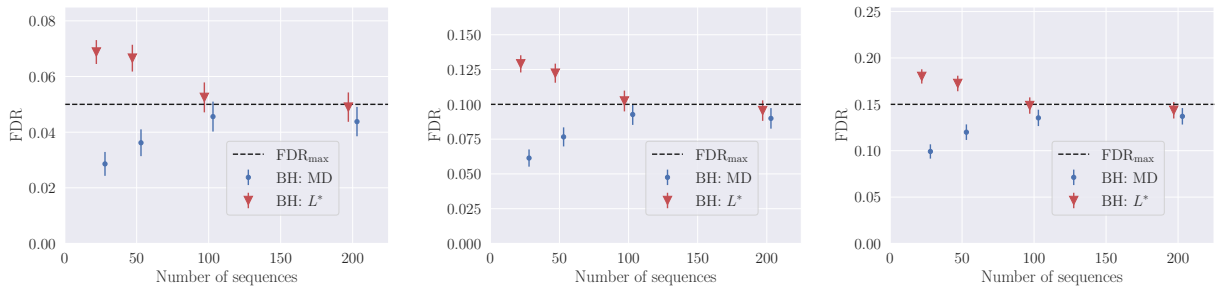
of these simulations suggest that the marginal model procedure using MD-corrected p-values should be preferred, as its use leads to better control of the FDR in comparison to L^*, while simultaneously giving either comparable or improved TPR values.

5. Application to Real Student Data

5.1. Affect Dynamics Data Sets

Our next analysis evaluates the performance of the marginal model procedure on actual student data. Specifically, we apply the technique to two different data sets consisting of affect sequences. Our first data set comes from students working in the Physics Playground learning environment (Shute and Ventura, 2013). The Baker-Rodrigo-Ocumpaugh Monitoring Protocol (BROMP) (Ocumpaugh et al., 2015) was used to record the affective states of 179 high school students working within this environment (Andres et al., 2015). For our purposes, we are interested in the states flow (FLO), confusion (CON), frustration (FRU), and boredom (BOR); the remaining states have all been merged into the dummy state NA. The recorded sequences for these students are relatively long, with the mean and median lengths being 135.2 and 126.0, respectively, with a standard deviation of 68.9; the minimum sequence length is 47, while the maximum is 272.

In contrast to the Physics Playground data, our second data set has very different characteristics. Namely, the sequences are much shorter, which makes for an interesting analysis, as we can see how the biases that have been observed in the simulated data affect the results from actual student data. This particular data set consists of sequences from 782 students working in the ASSISTments platform (Heffernan and Heffernan, 2014), with BROMP again being used to record the student affective states (Botelho et al., 2017); as before, we focus on the states flow (FLO), confusion (CON), frustration (FRU), and boredom (BOR), with any remaining states being merged into the dummy state NA. The mean and median lengths of the sequences in this data set are 9.6 and 9.0, respectively, with a standard deviation of 5.2. The minimum sequence



(a) $q = 0.05$ (b) $q = 0.1$ (c) $q = 0.15$

Figure 9: Comparison of the estimated FDR using a value of $\gamma = 0$; the BH procedure; and either L^* or the marginal model with MD-corrected p-values. Vertical lines represent the 99% confidence interval for each estimated FDR value. To avoid overlapping points, the plotted x-values for L^* and the marginal model have been shifted by -3 and $+3$, respectively.

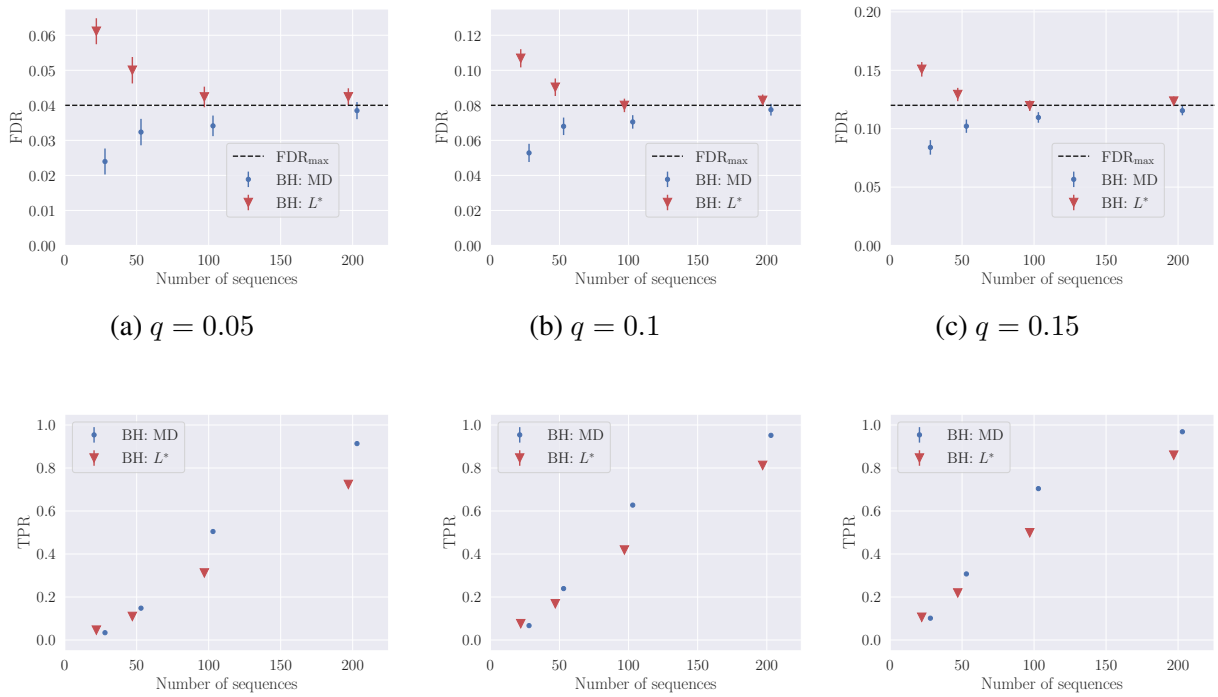
length is 3, while the maximum is 37.

5.2. Results: Marginal Model and the L Statistic

Based on the experiments from Bosch and Paquette (2021), as well as our results in Section 2.4, we expect the relatively long sequence lengths in the Physics Playground data to minimize the bias in the L statistic values. In comparison, we expect to see some evidence of this bias in the ASSISTments data, due to the very short sequence lengths. The results from applying the marginal model, as well as the corresponding L values, are shown in Tables 5 and 6. For the marginal model we use an exchangeable correlation structure as, overall, it gives better performance in comparison to the autoregressive structure.[8]

To adjust for the number of statistical tests being performed, we have highlighted—in bold—the transition pairs that are statistically significant after applying the Benjamini-Hochberg (BH) procedure with a value of $q = 0.1$. Regarding this choice of q, we first note that the results from our simulations in the previous section showed the combination of the BH procedure and MD-corrected p-values controlled the FDR at or below the expected level in all our experiments. Additionally, while a threshold value of $q = 0.05$ is commonly used with the BH procedure, arguments have been made that this is too low for many situations—for example, if the analysis is more exploratory in nature, or if follow up studies are relatively inexpensive—and that this choice of value may be due to confusion between the FDR and the false positive rate (James et al., 2021; McDonald, 2014). Thus, for these reasons we choose our value of $q = 0.1$, and we submit that this is a reasonable threshold for analyzing affect dynamics data. Lastly, we note that the focus in this analysis is more on comparing the marginal model and L statistic results, and less about identifying and interpreting significant affect transitions. To that end, and to more faithfully simulate the results from an actual study of affect transitions, we perform the

[8] While many applications of the autoregressive structure deal with time scales on the order of weeks, months, or years—e.g., epidemiological studies—the time scales for our data sets are much smaller, on the order of minutes or hours. Thus, due to these small time scales it's plausible that the dependence in our data is relatively constant over time, thereby making the exchangeable structure a better fit.



(a) $q = 0.05$ (b) $q = 0.1$ (c) $q = 0.15$

(d) $q = 0.05$ (e) $q = 0.1$ (f) $q = 0.15$

Figure 10: Comparison of the estimated FDR—(a) through (c)—and TPR—(d) through (f)—using a value of $\gamma = 0.05$; the BH procedure; and either L^* or the marginal model with MD-corrected *p*-values. Vertical lines represent the 99% confidence interval for each estimated FDR value (confidence intervals for the estimated TPR values are too small to be shown, as they are completely hidden behind the markers in all cases). To avoid overlapping points, the plotted x-values for L^* and the marginal model have been shifted by -3 and $+3$, respectively.

BH procedure twice per data set: once for all the marginal model values, and then separately for all the L statistic values.

Starting with the results from the Physics Playground data, it appears that the longer sequences in this data set have mitigated the effects of the bias with the L values. That is, there are only four transition pairs in the Physics Playground data for which the sign of the L statistic differs from the sign of the corresponding β_1 value; in each of these examples, the confidence interval for β_1 is relatively wide and contains zero, indicating there is a fair amount of uncertainty with the sign of the parameter. Furthermore, all of the self-transitions values, for both β_1 and L, are positive, with $p \ll 0.001$ in all cases; note that while the bias with the L statistic can heavily skew the estimates for self-transitions in the negative direction—e.g., see Figure 1—this does not appear to be the case here, most likely because of the long sequences of transitions.

Next, looking at the results for the ASSISTments data, there appears to be evidence of the bias in the L values for these shorter sequences. For example, recall that on short sequences of simulated data, the L statistic returns considerably lower than expected values for self-transitions. Thus, it is instructive to see that in all self-transition cases the L values are negative and significantly different from zero, while four of the five corresponding β_1 values

prev	*next*	All transitions: Physics Playground (longer sequences)						
		Marginal model					L statistic	
		β_1	p-value	95% CI	$x = 0$	$x = 1$	Mean	p-value
FLO	FLO	**0.78**	0.000	(0.67, 0.89)	0.61	0.78	**0.13**	0.000
	CON	**-0.47**	0.000	(-0.63, -0.32)	0.08	0.05	**-0.01**	0.000
	FRU	**-0.63**	0.000	(-0.81, -0.45)	0.08	0.04	**-0.01**	0.000
	BOR	**-1.43**	0.000	(-1.72, -1.14)	0.06	0.02	**-0.02**	0.000
	NA	**-0.37**	0.000	(-0.48, -0.25)	0.15	0.11	**-0.01**	0.000
CON	FLO	**-0.62**	0.000	(-0.76, -0.48)	0.74	0.60	**-0.71**	0.003
	CON	**1.33**	0.000	(1.10, 1.57)	0.05	0.18	**0.09**	0.000
	FRU	**0.38**	0.001	(0.16, 0.59)	0.05	0.07	**0.04**	0.013
	BOR	**-0.78**	0.068	(-1.61, 0.06)	0.03	0.01	**-0.02**	0.014
	NA	-0.06	0.468	(-0.21, 0.10)	0.13	0.12	0.01	0.700
FRU	FLO	**-0.81**	0.000	(-0.98, -0.64)	0.74	0.56	**-0.42**	0.001
	CON	0.14	0.366	(-0.17, 0.45)	0.06	0.07	-0.00	0.983
	FRU	**1.60**	0.000	(1.35, 1.85)	0.04	0.18	**0.07**	0.000
	BOR	**0.38**	0.037	(0.02, 0.75)	0.03	0.04	**0.03**	0.018
	NA	-0.00	0.979	(-0.19, 0.18)	0.12	0.12	0.03	0.163
BOR	FLO	**-1.28**	0.000	(-1.53, -1.03)	0.74	0.44	**-0.77**	0.001
	CON	**-1.16**	0.001	(-1.83, -0.49)	0.06	0.02	**-0.05**	0.000
	FRU	-0.17	0.296	(-0.48, 0.15)	0.05	0.04	0.01	0.390
	BOR	**2.96**	0.000	(2.46, 3.46)	0.02	0.27	**0.23**	0.000
	NA	-0.25	0.142	(-0.58, 0.08)	0.13	0.10	-0.02	0.400
NA	FLO	**-0.17**	0.008	(-0.29, -0.04)	0.73	0.70	-0.01	0.804
	CON	**-0.25**	0.004	(-0.42, -0.08)	0.06	0.05	**-0.01**	0.076
	FRU	**-0.47**	0.000	(-0.65, -0.28)	0.05	0.03	**-0.02**	0.010
	BOR	**-0.78**	0.003	(-1.31, -0.26)	0.03	0.02	**-0.02**	0.001
	NA	**0.65**	0.000	(0.51, 0.80)	0.11	0.20	**0.06**	0.000

Table 5: Comparison of the marginal model and L statistic on the Physics Playground (Andres et al., 2015) data set. Bold values are statistically significant after applying the Benjamini-Hochberg procedure with a value of $q = 0.1$. Note that, when applying the Benjamini-Hochberg procedure, we have applied it separately for the marginal model values and the L values, in order to simulate the workflow in a typical study of affect transitions.

prev	*next*	All transitions: ASSISTments (shorter sequences)						
		Marginal model					L statistic	
		β_1	p-value	95% CI	$x = 0$	$x = 1$	Mean	p-value
FLO	FLO	**0.51**	0.000	(0.39, 0.64)	0.57	0.69	**-0.09**	0.000
	CON	**-0.35**	0.002	(-0.58, -0.13)	0.08	0.06	**0.02**	0.001
	FRU	-0.21	0.171	(-0.52, 0.09)	0.04	0.03	0.00	0.285
	BOR	**-0.38**	0.000	(-0.56, -0.20)	0.12	0.08	**0.03**	0.002
	NA	**-0.61**	0.000	(-0.77, -0.46)	0.22	0.13	-0.01	0.293
CON	FLO	-0.02	0.871	(-0.26, 0.22)	0.64	0.64	**0.18**	0.004
	CON	**0.77**	0.000	(0.38, 1.16)	0.06	0.12	**-0.12**	0.000
	FRU	0.35	0.273	(-0.28, 0.98)	0.03	0.04	0.00	0.797
	BOR	-0.12	0.502	(-0.49, 0.24)	0.10	0.09	-0.02	0.339
	NA	-0.02	0.879	(-0.31, 0.27)	0.16	0.16	0.03	0.394
FRU	FLO	**-0.48**	0.000	(-0.75, -0.21)	0.64	0.53	-0.07	0.401
	CON	0.48	0.071	(-0.04, 1.00)	0.06	0.10	0.04	0.163
	FRU	0.04	0.961	(-1.42, 1.49)	0.03	0.03	**-0.10**	0.000
	BOR	**0.52**	0.012	(0.12, 0.93)	0.10	0.15	0.05	0.233
	NA	**0.40**	0.021	(0.06, 0.75)	0.16	0.22	**0.09**	0.032
BOR	FLO	**-0.28**	0.004	(-0.47, -0.09)	0.65	0.58	**0.18**	0.001
	CON	-0.31	0.138	(-0.71, 0.10)	0.06	0.05	-0.01	0.314
	FRU	0.31	0.231	(-0.20, 0.81)	0.03	0.04	0.01	0.280
	BOR	**0.73**	0.000	(0.39, 1.07)	0.09	0.16	**-0.09**	0.000
	NA	0.07	0.555	(-0.17, 0.31)	0.16	0.17	0.01	0.625
NA	FLO	**-0.37**	0.000	(-0.52, -0.23)	0.65	0.57	**0.11**	0.033
	CON	0.07	0.596	(-0.20, 0.34)	0.06	0.07	0.01	0.434
	FRU	-0.06	0.780	(-0.49, 0.37)	0.03	0.03	0.00	0.771
	BOR	-0.07	0.528	(-0.28, 0.15)	0.10	0.09	0.02	0.122
	NA	**0.79**	0.000	(0.59, 0.99)	0.14	0.26	**-0.05**	0.002

Table 6: Comparison of the marginal model and L statistic on the ASSISTments (Botelho et al., 2017) data set. Bold values are statistically significant after applying the Benjamini-Hochberg procedure with a value of $q = 0.1$. Note that, when applying the Benjamini-Hochberg procedure, we have applied it separately for the marginal model values and the L values, in order to simulate the workflow in a typical study of affect transitions.

are positive and significantly different from zero. For comparison, of the 20 transitions between different states, there are only 4 transition pairs that have negative L values, none of which are significantly different from zero; thus the positive bias when the L statistic is applied to transitions between different states seems to be a factor. Furthermore, in all four of these cases with negative L values, the corresponding β_1 values are also negative, possibly indicating that the negative relationships between the states are strong enough to overcome the positive bias with the L statistic. Thus, the overall results on the ASSISTments data set are seemingly consistent with the biases that appear in the experiments on simulated data.

5.3. Results: No Self-Transitions

In this section we analyze the same data sets with the influence of self-transitions removed. For this analysis we compare the results from using the marginal model procedure with those from an application of L^*. Starting with the Physics Playground data set, in Table 7 we can see that in all but one row the signs of β_1 and L^* agree—in the one case they don't agree, transitions from BOR to FLO, there is a large amount of uncertainty associated with the sign of the L^* value.

Next, in Table 8 we show the results on the ASSISTments data set with the influence of self-transitions removed. Here, there is a lot more uncertainty in all of the estimates—this can be seen, for example, in the generally large confidence intervals for β_1 that contain zero in the majority of cases. Thus, it's harder to make a direct comparison of the procedures on this data set once self-transitions have been removed. However, it's worth comparing these results with those in Table 6, where self-transitions have not been removed from the ASSISTments data. Specifically, the marginal model estimates in Table 6 are in general more precise, which can be partly explained by the fact that removing self-transitions leaves less data for the model to construct its estimates with. That is, while the full data set contains 6,755 transitions, after removing self-transitions the number remaining are as follows: 2,325 (FLO); 6,344 (CON); 6,532 (FRU); 6,164 (BOR); 5,655 (NA). The difference is particularly large for transitions starting in FLO, as overall this is the most common state in the ASSISTments data.

6. Discussion

In this work we have attempted to give a comprehensive examination of several issues related to the analysis of transitions in sequential data. In our first analysis, we addressed the problem discussed in Bosch and Paquette (2021), where it was shown that several commonly used transition metrics suffer from a bias that can inflate the importance of transition pairs. After replicating the numerical experiments in Bosch and Paquette (2021), we next presented a theoretical explanation for the underlying cause of this bias, where we argued that it's a consequence of the averaging procedure commonly used in the computations of transition metrics. Based on these results, we then outlined a procedure for measuring the importance and significance of transition pairs. This procedure takes the form of a logistic regression that estimates the probability of transitioning to a state, depending on the occurrence of a previous state; the parameters for the regression were obtained using marginal models. To show that this procedure does not suffer from the bias inherent in other transition metrics, we examined its effectiveness on simulated data.

Our second analysis looked at the special case of removing the influence of self-transitions. As it was shown in Karumbaiah et al. (2019) that excluding self-transitions has unintended

prev	*next*	No self-transitions: Physics Playground (longer sequences)						
		Marginal model					L^*	
		β_1	p-value	95% CI	$x = 0$	$x = 1$	Mean	p-value
FLO	FLO	–	–	–	–	–	–	–
	CON	**0.14**	0.036	(0.01, 0.27)	0.21	0.24	**0.02**	0.033
	FRU	-0.06	0.444	(-0.20, 0.09)	0.20	0.19	**-0.03**	0.009
	BOR	**-0.69**	0.000	(-0.87, -0.51)	0.13	0.07	**-0.05**	0.000
	NA	**0.24**	0.000	(0.12, 0.36)	0.45	0.51	0.02	0.361
CON	FLO	**-0.22**	0.008	(-0.39, -0.06)	0.78	0.74	-0.70	0.098
	CON	–	–	–	–	–	–	–
	FRU	**0.59**	0.000	(0.38, 0.80)	0.05	0.09	**0.05**	0.003
	BOR	-0.42	0.218	(-1.09, 0.25)	0.03	0.02	-0.01	0.088
	NA	0.13	0.133	(-0.04, 0.30)	0.13	0.15	0.02	0.345
FRU	FLO	**-0.41**	0.000	(-0.55, -0.26)	0.77	0.69	**-0.32**	0.017
	CON	**0.42**	0.007	(0.11, 0.72)	0.07	0.10	0.01	0.220
	FRU	–	–	–	–	–	–	–
	BOR	**0.62**	0.001	(0.27, 0.97)	0.03	0.06	**0.04**	0.007
	NA	0.15	0.152	(-0.05, 0.35)	0.13	0.15	0.04	0.074
BOR	FLO	-0.21	0.134	(-0.49, 0.06)	0.75	0.71	0.01	0.961
	CON	-0.49	0.105	(-1.09, 0.10)	0.07	0.04	**-0.04**	0.002
	FRU	**0.39**	0.009	(0.10, 0.68)	0.05	0.08	0.03	0.143
	BOR	–	–	–	–	–	–	–
	NA	**0.38**	0.056	(-0.01, 0.78)	0.13	0.18	0.03	0.347
NA	FLO	**0.39**	0.000	(0.24, 0.54)	0.83	0.88	**0.26**	0.001
	CON	-0.15	0.088	(-0.33, 0.02)	0.07	0.06	-0.01	0.192
	FRU	**-0.41**	0.000	(-0.62, -0.21)	0.06	0.04	**-0.02**	0.011
	BOR	**-0.65**	0.015	(-1.17, -0.13)	0.04	0.02	**-0.01**	0.012
	NA	–	–	–	–	–	–	–

Table 7: Comparison of the marginal model and L^* on the Physics Playground (Andres et al., 2015) data set with self-transitions removed. Bold values are statistically significant after applying the Benjamini-Hochberg procedure with a value of $q = 0.1$. Note that, when applying the Benjamini-Hochberg procedure, we have applied it separately for the marginal model values and the L^* values, in order to simulate the workflow in a typical study of affect transitions.

prev	*next*	No self-transitions: ASSISTments (shorter sequences)						
		Marginal model					L^*	
		β_1	p-value	95% CI	$x = 0$	$x = 1$	Mean	p-value
FLO	FLO	–	–	–	–	–	–	–
	CON	0.20	0.055	(-0.00, 0.40)	0.17	0.20	0.02	0.213
	FRU	0.20	0.161	(-0.08, 0.49)	0.08	0.10	0.01	0.323
	BOR	-0.06	0.483	(-0.23, 0.11)	0.26	0.25	0.01	0.793
	NA	-0.13	0.106	(-0.28, 0.03)	0.49	0.46	-0.06	0.048
CON	FLO	-0.00	0.980	(-0.25, 0.25)	0.68	0.68	-0.04	0.721
	CON	–	–	–	–	–	–	–
	FRU	0.44	0.182	(-0.21, 1.09)	0.03	0.05	0.00	0.938
	BOR	-0.16	0.468	(-0.60, 0.27)	0.11	0.09	-0.05	0.101
	NA	0.02	0.886	(-0.29, 0.33)	0.17	0.18	-0.02	0.667
FRU	FLO	**-0.64**	0.000	(-0.93, -0.35)	0.66	0.51	-0.29	0.036
	CON	0.53	0.053	(-0.01, 1.07)	0.07	0.11	0.03	0.372
	FRU	–	–	–	–	–	–	–
	BOR	0.49	0.037	(0.03, 0.94)	0.10	0.16	0.03	0.551
	NA	0.42	0.023	(0.06, 0.78)	0.17	0.23	0.05	0.216
BOR	FLO	-0.21	0.074	(-0.43, 0.02)	0.71	0.67	0.03	0.712
	CON	-0.09	0.669	(-0.53, 0.34)	0.07	0.07	-0.02	0.317
	FRU	0.54	0.038	(0.03, 1.04)	0.04	0.06	0.01	0.459
	BOR	–	–	–	–	–	–	–
	NA	0.23	0.100	(-0.04, 0.50)	0.18	0.21	-0.01	0.766
NA	FLO	-0.15	0.111	(-0.33, 0.03)	0.76	0.73	-0.04	0.694
	CON	0.34	0.016	(0.06, 0.62)	0.07	0.10	0.00	0.767
	FRU	0.15	0.482	(-0.27, 0.58)	0.04	0.05	-0.00	0.989
	BOR	0.04	0.765	(-0.21, 0.29)	0.12	0.12	0.00	0.885
	NA	–	–	–	–	–	–	–

Table 8: Comparison of the marginal model and L^* on the ASSISTments (Botelho et al., 2017) data set with self-transitions removed. Bold values are statistically significant after applying the Benjamini-Hochberg procedure with a value of $q = 0.1$. Note that, when applying the Benjamini-Hochberg procedure, we have applied it separately for the marginal model values and the L^* values, in order to simulate the workflow in a typical study of affect transitions.

consequences when used in conjunction with the L statistic, many recent works have proposed fixes for this issue (Bosch and Paquette, 2021; Karumbaiah et al., 2019; Karumbaiah et al., 2021; Matayoshi and Karumbaiah, 2020). However, in the interest of having a uniform procedure that can be used whether or not a researcher wants to remove the influence of self-transitions, we proposed an extension of the marginal model procedure to this specific situation. After giving theoretical arguments for why the proposed procedure is valid in this situation, we then evaluated its performance on simulated data to show that it does not suffer from the aforementioned issues.

We next discussed the problem of controlling for multiple comparisons when analyzing transitions in sequential data. In particular, the results in Matayoshi and Karumbaiah (2021a) suggested that there are issues with applying the Benjamini-Hochberg (BH) procedure to the analysis of sequential data, with their experiments showing this leads to an inflated rate of false discoveries in many situations. Thus, to address this issue we borrowed a technique from the biostatistics literature that adjusts the standard errors and p-values resulting from our estimates of the marginal model regression coefficients. We then evaluated this corrected procedure on simulated data, where we saw that, after applying these adjustments, the BH procedure gave good control of the false discovery rate (FDR). Additionally, in the case when self-transitions are removed, we compared the marginal model procedure to the performance of L^*, a statistic specifically developed for use when one wants to remove the influence of self-transitions. We saw in all cases that using the marginal model procedure led to better control of the FDR in comparison to L^*, with the differences being especially notable with smaller numbers of sequences. Furthermore, a comparison of the true positive rates (TPR) revealed that, in most cases, the marginal model is also better at identifying the true discoveries, sometimes substantially so. Thus, these results presented strong evidence favoring the use of the marginal model procedure over L^*.

Finally, we concluded by synthesizing all of these results with an analysis on real student data. Here, we saw further evidence that the marginal model procedure does not suffer from the bias inherent in other transition metrics. Additionally, informed by our simulations pertaining to the problem of multiple comparisons, we applied the BH procedure with a threshold value of $q = 0.1$ when analyzing the student data sets. While we believe that, in many cases, this is a reasonable threshold value, it should be mentioned that the specific choice of value could very much depend on the nature of the analysis. For example, if the research is exploratory in nature, with the goal being to identify transition pairs that might be worth further study, higher threshold values such as 0.15 or 0.2 might be appropriate.

Additionally, we encourage researchers who are analyzing transitions in sequential data to think critically about the aims of their study and not to consider the "statistical significance" results as a final outcome. Specifically, in our analyses in Section 5 we've included other results from the marginal model procedure, such as confidence intervals and the state transition probability estimates. In fact, we believe the probability estimates are one of the strengths of this approach. To illustrate this point, consider the results from the Physics Playground data set in Table 5 where, after applying the BH procedure, 20 of the 25 transition pairs are considered statistically significant—such a result is arguably of somewhat limited value, as it's not immediately clear which of these 20 transitions are most important. However, by comparing the probability estimates for each of these transition pairs when $x = 0$ and $x = 1$, we can see that the practical significance of the pairs varies greatly. For example, the estimated probability of a transition from BOR to FLO is 0.44; however, starting in any state other than BOR the estimated probability is 0.74, a substantial increase. Compare this with the probability estimate for

transitions from BOR to CON, which is 0.06, and which then decreases to 0.02 when starting in any state other than BOR. While this difference is considered statistically significant, from a practical standpoint it seems less important and useful in comparison to the large difference that occurs with transitions from BOR to FLO.

Finally, we note that the approach outlined here is flexible, as it can be applied to estimate and measure the effects of other relationships beyond a single transition between states. As an example, suppose we are interested in whether starting in state A has an influence on the appearance of the sequence BAB as the next three states. In this case, we simply need to change our response variable to fit the situation. Rather than defining the response variable based on the next state, we simply change the definition so that it has a value of one if the next three states are BAB, and a value of zero otherwise. Or, perhaps we are still interested in a transition to a single state, but rather than looking at the starting states individually, we would rather directly compare their influence simultaneously. In this case, and assuming we don't want to remove the influence of self-transitions, we can use different indicator variables for the starting states, and then compare the coefficients of these indicator variables to get a relative ordering of the importance of the different starting states. Note, however, there are complications with this last approach when the researcher would like to remove the influence of self-transitions, as the procedure outlined in Section 3.1 is not directly applicable. Thus, we are currently looking into this situation in more detail.

REFERENCES

AKAIKE, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19,* 6, 716–723.

ANDRADE, A., DANISH, J., AND MALTESE, A. 2017. A measurement model of gestures in an embodied learning environment: Accounting for temporal dependencies. *Journal of learning Analytics 4*, 18–46.

ANDRES, J. M. L., RODRIGO, M. M. T., SUGAY, J. O., BANAWAN, M. P., PAREDES, Y. V. M., CRUZ, J. S. D., AND PALAOAG, T. D. 2015. More fun in the Philippines? Factors affecting transfer of western field methods to one developing world context. In *Proceedings of the Sixth International Workshop on Culturally-Aware Tutoring Systems at the 17th International Conference on Artificial Intelligence in Education*, J. Boticario and K. Muldner, Eds. CEUR Workshop Proceedings, vol. 1432. 31–40.

ANGRIST, J. D. AND PISCHKE, J.-S. 2008. *Mostly Harmless Econometrics*. Princeton University Press.

BENJAMINI, Y. 2010. Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72,* 4, 405–416.

BENJAMINI, Y. AND HOCHBERG, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological) 57,* 1, 289–300.

BENJAMINI, Y. AND YEKUTIELI, D. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics 29,* 4, 1165–1188.

BISWAS, G., JEONG, H., KINNEBREW, J., SULCER, B., AND ROSCOE, R. D. 2010. Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology Enhanced Learning 5,* 2, 123–152.

BOSCH, N. AND D'MELLO, S. 2017. The affective experience of novice computer programmers. *International Journal of Artificial Intelligence in Education 27,* 1, 181–206.

BOSCH, N. AND PAQUETTE, L. 2021. What s next? Sequence length and impossible loops in state transition measurement. *Journal of Educational Data Mining 13,* 1, 1–23.

BOTELHO, A. F., BAKER, R. S., AND HEFFERNAN, N. T. 2017. Improving sensor-free affect detection using deep learning. In *Proceedings of the 18th International Conference on Artificial Intelligence in Education*, E. André, R. Baker, X. Hu, M. M. T. Rodrigo, and B. du Boulay, Eds. Springer International Publishing, Cham, 40–51.

D'MELLO, S. AND GRAESSER, A. 2012. Dynamics of affective states during complex learning. *Learning and Instruction 22,* 2, 145–157.

D'MELLO, S., TAYLOR, R. S., AND GRAESSER, A. 2007. Monitoring affective trajectories during complex learning. In *Proceedings of the 29th Annual Cognitive Science Society*, D. S. McNamara and J. G. Trafton, Eds. Cognitive Science Society, Austin, TX, 203–208.

FARCOMENI, A. 2006. More powerful control of the false discovery rate under dependence. *Statistical Methods and Applications 15,* 1, 43–73.

GILOVICH, T., VALLONE, R., AND TVERSKY, A. 1985. The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology 17,* 3, 295–314.

GOEMAN, J. J. AND SOLARI, A. 2014. Multiple hypothesis testing in genomics. *Statistics in medicine 33,* 11, 1946–1978.

HARDIN, J. W. AND HILBE, J. M. 2012. *Generalized Estimating Equations*. Chapman and Hall/CRC.

HEAGERTY, P. J. AND ZEGER, S. L. 2000. Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statistical Science 15,* 1, 1–26.

HEFFERNAN, N. T. AND HEFFERNAN, C. L. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education 24,* 4, 470–497.

JAMES, G., WITTEN, D., HASTIE, T., AND TIBSHIRANI, R. 2021. *An Introduction to Statistical Learning*, Second ed. Springer.

KARUMBAIAH, S., ANDRES, J. M. A. L., ANTHONY., F., BOTELHO, BAKER, R., AND OCUMPAUGH, J. L. 2018. The implications of a subtle difference in the calculation of affect dynamics. In *Proceedings of the 26th International Conference on Computers in Education*, J. C. Yang, M. Chang, L.-H. Wong, and M. M. T. Rodrigo, Eds. 29–38.

KARUMBAIAH, S., BAKER, R., OCUMPAUGH, J., AND ANDRES, A. 2021. A re-analysis and synthesis of data on affect dynamics in learning. *IEEE Transactions on Affective Computing*.

KARUMBAIAH, S., BAKER, R. S., AND OCUMPAUGH, J. 2019. The case of self-transitions in affective dynamics. In *Proceedings of the 20th International Conference on Artificial Intelligence in Education*, S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, and R. Luckin, Eds. Springer International Publishing, Cham, 172–181.

KIM, K. I. AND VAN DE WIEL, M. A. 2008. Effects of dependence in high-dimensional multiple testing problems. *BMC bioinformatics 9,* 1, 1–12.

KNIGHT, S., WISE, A. F., AND CHEN, B. 2017. Time for change: Why learning analytics needs temporal analysis. *Journal of Learning Analytics 4,* 3, 7–17.

LI, P. AND REDDEN, D. T. 2015. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Statistics in Medicine 34,* 2, 281–296.

LIANG, K.-Y. AND ZEGER, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika 73,* 1, 13–22.

MACKINNON, J. G. AND WHITE, H. 1985. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics 29,* 3, 305–325.

MAHZOON, M. J., MAHER, M. L., ELTAYEBY, O., DOU, W., AND GRACE, K. 2018. A sequence data model for analyzing temporal patterns of student data. *Journal of Learning Analytics 5,* 1, 55–74.

MANCL, L. A. AND DEROUEN, T. A. 2001. A covariance estimator for GEE with improved small-sample properties. *Biometrics 57,* 1, 126–134.

MATAYOSHI, J. AND KARUMBAIAH, S. 2020. Adjusting the L statistic when self-transitions are excluded in affect dynamics. *Journal of Educational Data Mining 12,* 4 (Dec.), 1–23.

MATAYOSHI, J. AND KARUMBAIAH, S. 2021a. Investigating the validity of methods used to adjust for multiple comparisons in educational data mining. In *Proceedings of the 14th International Conference on Educational Data Mining*. 33–45.

MATAYOSHI, J. AND KARUMBAIAH, S. 2021b. Using marginal models to adjust for statistical bias in the analysis of state transitions. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. 449– 455.

MCDONALD, J. 2014. *Handbook of Biological Statistics (3rd ed.)*. Sparky House Publishing.

MILLER, J. B. AND SANJURJO, A. 2018. Surprised by the gamblers and hot hand fallacies? A truth in the law of small numbers. *Econometrica* 6, 2019–2047.

OCUMPAUGH, J., BAKER, R. S., AND RODRIGO, M. M. T. 2015. Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 technical and training manual. *New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences 60.*

PAN, W. 2001. Akaike's information criterion in generalized estimating equations. *Biometrics 57,* 1, 120–125.

REINER-BENAIM, A. 2007. FDR control by the BH procedure for two-sided correlated tests with implications to gene expression data analysis. *Biometrical Journal 49,* 1, 107–126.

REINER-BENAIM, A., YEKUTIELI, D., LETWIN, N. E., ELMER, G. I., LEE, N. H., KAFKAFI, N., AND BENJAMINI, Y. 2007. Associating quantitative behavioral traits with gene expression in the brain: Searching for diamonds in the hay. *Bioinformatics 23,* 17, 2239–2246.

SACKETT, G. P. 1979. The lag sequential analysis of contingency and cyclicity in behavioral interaction research. *Handbook of Infant Development 1*, 623–649.

SEABOLD, S. AND PERKTOLD, J. 2010. Statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*.

SHUTE, V. J. AND VENTURA, M. 2013. *Stealth assessment: Measuring and supporting learning in video games*. MIT Press.

SNIJDERS, T. A. AND BOSKER, R. J. 2012. *Multilevel Analysis: An introduction to Basic and Advanced Multilevel Modeling*. Sage.

STOREY, J. D. AND TIBSHIRANI, R. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences 100,* 16, 9440–9445.

SZMARAGD, C., CLARKE, P., AND STEELE, F. 2013. Subject specific and population average models for binary longitudinal data: a tutorial. *Longitudinal and Life Course Studies 4,* 2, 147–165.

WILLIAMS, V. S., JONES, L. V., AND TUKEY, J. W. 1999. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of educational and behavioral statistics 24,* 1, 42–69.

YEKUTIELI, D. 2008. False discovery rate control for non-positively regression dependent test statistics. *Journal of Statistical Planning and Inference 138,* 2, 405–415.