Investigating the Validity of Methods Used to Adjust for Multiple Comparisons in Educational Data Mining

Jeffrey Matayoshi
McGraw Hill ALEKS
Irvine, CA, USA
jeffrey.matayoshi@aleks.com

Shamya Karumbaiah
University of Pennsylvania
Philadelphia, Pennsylvania, USA
shamya@upenn.com

ABSTRACT

Research studies in Educational Data Mining (EDM) often involve several variables related to student learning activities. As such, it may be necessary to run multiple statistical tests simultaneously, thereby leading to the problem of multiple comparisons. The Benjamini-Hochberg (BH) procedure is commonly used in EDM research to address this issue, and it has proven to be a useful method. However, the main limitation of the procedure is that it requires the statistical tests to either be independent or satisfy certain dependency conditions. The Benjamini-Yekutieli (BY) procedure is an alternative that can be applied under arbitrary dependence assumptions, but this extra flexibility comes with a loss of statistical power; hence, the BH procedure is preferred whenever it can be properly applied. Based on these considerations, in this work we employ simulation studies to assess the validity of the BH procedure in two scenarios common to EDM research. The first scenario considers the evaluation and comparison of different classification models—such an analysis might occur, for instance, during the model tuning and validation stage of a study. Then, in the second scenario we look at experiments involving the study of state transitions in sequential data, examples of which occur in affect dynamics research. We find that the BH procedure performs as expected when used with simulated classification model predictions; however, when applied to simulated sequential data, it does not perform at the expected level. Based on these results, as well as previous studies evaluating the BH and BY methods, we discuss the appropriate usage of these procedures for the scenarios under examination.

Keywords

Multiple comparisons, false discovery rate, Benjamini-Hochberg, Benjamini-Yekutieli

1. INTRODUCTION

Consider a statistical analysis that tests several different null hypotheses, either on a single data set, or on related data sets. In such a scenario, the probability of making a *discovery*—i.e., rejecting a null hypothesis—is higher than in an analysis involving a single null hypothesis. 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.

Studies in Educational Data Mining (EDM) and related fields are shaping the ongoing research and development of learning systems that are increasingly becoming part of everyday classrooms—thus directly impacting student lives. Greater attention is needed to ensure that the conclusions drawn from these studies are reliable. Along these lines, controlling for multiple comparisons is an important consideration, as it has been argued that addressing the issue is a major factor in ensuring the replicability of scientific results [2]. Additionally, many exaggerated or even incorrect results can be explained by the testing of multiple hypotheses without adjusting for the number of comparisons [34, 40]; while this issue commonly occurs with observational data, experimental studies are not immune to the problem [30].

The main focus of this study is the Benjamini-Hochberg (BH) procedure [3], a method that is commonly applied in EDM research to control the *false discovery rate* (FDR)—defined as the expected rate of false discoveries among all the discoveries made—when multiple statistical tests are used. 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 [3, 4]. The Benjamini-Yekutieli (BY) procedure is an alternative method that can be used under arbitrary dependence assumptions among the statistical tests [4]. As the BY procedure is more generally applicable than the BH procedure, it is by necessity more conservative and thus less likely to classify a result as being statistically significant; in turn, this causes it to have lower statistical power compared to the BH procedure. Thus, the BH procedure is to be preferred over the BY procedure whenever it can be properly applied.

However, the difficulty is that verifying the conditions for applying the BH procedure is not always straightforward; while some scenarios have been mathematically proven to satisfy these conditions, many common examples have not been. For instance, as of 2010 the case of pairwise comparisons had not been mathematically proven to satisfy the

conditions for using the BH procedure [1], and to the best of our knowledge that has not changed in the interim. Because it's not always clear if the conditions for applying the BH procedure are satisfied, it is often used without any theoretical guarantees on its performance [15]. In other situations, researchers may resort to using both the BH and BY procedures and comparing the results [28]. Motivated by these issues, in this work we investigate two different scenarios that occur within EDM research, with the goal of understanding if the BH procedure is appropriate for each situation. In both scenarios, we assume that a researcher wants to control the FDR, ideally with the BH procedure, but is unsure if it will work as desired. As we are unable to provide mathematical proofs for these scenarios, we instead turn to simulation studies, a procedure that is commonly used to investigate the performance of multiple comparison procedures [1, 3, 14, 22, 31, 32, 38, 39].

The outline of the paper is as follows. We first discuss the specifics of the BH and BY procedures and how to apply them when performing multiple hypothesis tests; additionally, we also look at how multiple comparisons are handled in the EDM community by surveying the literature from the last five EDM conference proceedings. Then, in the remainder of the paper we evaluate the BH and BY procedures for two scenarios that EDM researchers may encounter in their work. The first scenario concerns the usage of these procedures for evaluating and comparing the performance of classification models. In this scenario, we make pairwise comparisons of simulated classifiers, using both accuracy and the area under the receiver operating characteristic curve (AUROC) to evaluate their performance; such a situation can occur, for example, when trying to find the best performing combinations of model algorithms and hyperparameters.

The next scenario we look at is the analysis of state transitions in sequential data. In such an analysis, researchers typically run several hypothesis tests to try and determine the importance of the various transitions between states. Examples of this occur in affect dynamics research, where the BH procedure is commonly used [18, 29]. Here, we run analyses on simulated sequences of transitions using two different statistical measures, and we then apply the BH and BY procedures and compare the results. Finally, based on the results of our numerical experiments, as well as the existing literature on controlling the FDR, we discuss the usage of the BH and BY procedures in these scenarios.

2. CONTROLLING FOR MULTIPLE COMPARISONS

2.1 Benjamini-Hochberg and Benjamini-Yekutieli Procedures

In this study we focus on procedures for controlling the false discovery rate (FDR). The FDR was introduced in [3], and it has since found widespread use in many scientific fields including education research [38], genetics [31, 35], and medical studies [4]. If we let $\mathbf{V}$ be the number of false discoveries and $\mathbf{S}$ be the number of true discoveries, as done in [3] we can define the quantity $\mathbf{Q}$ as

$$\mathbf{Q} = \begin{cases} \frac{\mathbf{V}}{\mathbf{V}+\mathbf{S}}, & \text{if } \mathbf{V}+\mathbf{S} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

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 [16], 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,[1] or when the research is of a slightly more exploratory nature.

In addition to introducing the FDR to the scientific literature, the authors in [3] also outlined the 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 [4], 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 [4] introduced the BY procedure for situations in which the BH procedure is not valid, and they proved that the BY 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

[1] 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 [1].

between these various terms can be summarized as follows.

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

(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

Let q represent our chosen threshold—or, level—for controlling the FDR, and define $\text{FDR}_{\max} = \frac{m_0}{m}q$. If the statistical tests are independent, or if they satisfy certain dependency conditions, it was shown in [4] 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 [4] 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.

2.2 Applications in EDM Research

To understand how EDM research is controlling for multiple comparisons, we reviewed EDM conference proceedings from the last five years (2016–2020). We found that, among the 22 papers that reported controlling for multiple comparisons,[2] half used the Bonferroni correction and half used the BH procedure, with no studies using the BY procedure. Based on the method used to perform the comparisons, the studies can be partitioned as follows: group comparison (8), pairwise comparison (8; including pairwise model comparison), correlation (4), and regression analysis (2). The studies involving group comparisons used statistical methods such as the Mann-Whitney U test, chi-squared test, t-test, and ANOVA. The studies employing pairwise comparisons used methods such as the Kruskal-Wallis test, Mann-Whitney U test, McNemar's test, chi-squared test, and t-test. Overall, these 22 studies investigated diverse educational constructs in virtual learning environments including affect, student behavior in MOOCs, help-seeking, collaboration, and self-regulation.

The choice between the Bonferroni correction and the BH procedure varied in the studies, as the selection was not completely determined by the study methodology. For instance, an exploratory study used the more conservative Bonferroni method for a correlational analysis [61], while an experimental study with group comparisons used the less conservative BH procedure [46]. For EDM research, selecting between the Bonferroni correction and the BH procedure may not be universal and likely depends on the context of the study. As an example, consider that an analysis examining student demographic differences on an important educational construct—such as self-efficacy, affect, or learning—likely has fewer data samples from underrepresented minorities [20]. In such a case, penalizing the statistical power with a more conservative method like the Bonferroni correction may lead to missed opportunities for critical discoveries related to equity. On the other hand, contrast this with the evaluation of an expensive and large-scale educational technology intervention in a public school system; given the costs involved, both financially and otherwise, it could be argued that such an evaluation requires a more conservative approach to control for false discoveries.

More broadly, EDM research may not always involve large data sets. This is particularly true for educational constructs that require resource-intensive data collection procedures—

[2] See Section 8 for the full list of references.

e.g., training coders, gathering approvals, and conducting classroom studies. Hence, using the Bonferroni correction to control for multiple comparisons at the expense of losing statistical power may not always be affordable. In contrast, using the BH procedure in scenarios that violate its statistical assumptions may lead to invalid conclusions. Our review of EDM studies from the last five years also revealed that the field may not be taking advantage of the BY procedure, especially in scenarios where it is difficult to meet the assumptions of the BH procedure. These observations are what motivated us to investigate the use of the BH and BY procedures in research settings relevant to EDM.

3. METHODS

In this section we outline the general procedure we follow for our 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 for the two scenarios we consider, 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. For example, in our comparisons of simulated classification models, the performance of each model is controlled by a single parameter; thus, when this parameter differs for two models, the null hypothesis that the models perform equally well is false.

Another important detail is that, as we are focusing on two particular scenarios, we can generate simulated data specific to these scenarios. That is, for the model comparison experiments we simulate both the classifier predictions and the ground truth labels; then, for the state transition analysis we generate simulated sequences of states. By simulating the underlying data for each scenario, we are attempting to evaluate the BH and BY 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 and BY procedures 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 [26]; thus, we evaluate a range of values in our simulations. Based on the statistical significance results from our application of the BH and BY procedures, we can compute $\mathbf{Q}$, the proportion of false discoveries among all the discoveries made, using (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 2.1.

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.[3] Using the notation in (2), the false positive rate can be written as $\frac{\mathbf{V}}{\mathbf{V}+\mathbf{U}}$. In comparison, $\mathbf{Q}$ is computed with the formula $\frac{\mathbf{V}}{\mathbf{V}+\mathbf{S}}$, 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 (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. MODEL COMPARISONS

The first scenario we study concerns the comparison of several classification models on a fixed set of test or validation data. A common example of this occurs during the model building process, where it may be necessary to evaluate the performance of many different combinations of classification models and hyperparameters. In such a case, it can be helpful for the researcher to run statistical tests to more precisely quantify the differences in performance. To that end, we focus on the pairwise comparisons of the classifiers, where we assume that the classifiers could have different underlying algorithms—e.g., logistic regression vs. random forest—or the same algorithm with different hyperparameters. We evaluate each pair of classifiers by looking at both the accuracy and the area under the receiver operating characteristic curve (AUROC). To measure the possible difference between the accuracy values of the models, we use McNemar's test [13, 27]. When conducting pairwise comparisons of classifier accuracy on a fixed set of test data—as opposed to a procedure such as k-fold cross-validation, where the test set varies—using McNemar's test is recommended [10]; for these evaluations we use the implementation in the `statsmodels` [33] Python library. Then, to compare the AUROC values we use DeLong's test [9], a method developed to statistically test for differences in AUROC values; specifically, we apply the fast version of the algorithm outlined in [36].[4]

Our simulations use the following procedure. We assume that we are evaluating the performance of a binary classifier on a test set containing n data points; for these simulations we use n-values of 500, 1000, and 5000. For each value of n, we sample n numbers uniformly at random from 0.01 to 0.99; we refer to this set of numbers as $\mathbf{U}_n$. In each simulation run, the numbers in $\mathbf{U}_n$ are used to generate the labels for our data using the following procedure. Let i be an integer from 1 to n, and let $p_i \in \mathbf{U}_n$. With probability p_i we assign a label of 1 to y_i; otherwise, with probability $1-p_i$ it is then given a label of 0. Note that the set $\mathbf{U}_n$ is generated once for each value of n, and this same set is then used repeatedly for all of our simulation runs with a test set of size n.

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

[4] The code for our implementation of the algorithm in [36], as well as for running all of our experiments, is available at https://github.com/jmatayoshi/multiple-comparisons.

Table 1: Accuracy and AUROC values for an example simulation run using a test set of size $n = 1000$.

σ	0.1	0.1	0.1	0.5	1	2
Accuracy	0.733	0.724	0.732	0.706	0.651	0.606
AUROC	0.824	0.821	0.824	0.787	0.721	0.655

We next describe our procedure for simulating the classifier predictions. Let c_{ij} represent the predicted probability given by classifier j for the i-th data point in our test set. To generate c_{ij}, we begin by converting $p_i \in \mathbf{U}_n$ to a $\mathbf{z}$-score. Then, to add noise to the classifier's prediction we randomly sample a value, s_{ij}, from a normal distribution with mean 0 and standard deviation σ_j, add this to the $\mathbf{z}$-score, and then convert everything back to a probability; the resulting value is c_{ij}. The size of σ_j controls the performance of the classifier, with lower values giving predicted probabilities that are less noisy and more likely to align with the class labels. Let Φ denote the cumulative distribution function (CDF) of the standard normal distribution. Our procedure for generating the classifier predictions can be summarized as follows.

1. $z_i = \Phi^{-1}(p_i)$
2. Draw sample value s_{ij} from $\mathcal{N}(0, \sigma_j^2)$
3. $c_{ij} = \Phi(z_i + s_{ij})$

To get an idea of the effect of different values of σ on the performance of our simulated classifier predictions, in Table 1 we show the accuracy and AUROC values from one simulation run, using different values of σ and a test set size of $n = 1000$. The three classifiers with σ values of 0.1 have the best performance, with accuracy values from 0.72 to 0.73 and AUROC values around 0.82. The other classifiers, to varying degrees, perform worse, with the lowest accuracy and AUROC values at roughly 0.61 and 0.66, respectively. Our initial analysis simulates the pairwise comparison of six different classification models, where all the classifiers are assumed to perform equally; specifically, we use a value of $\sigma = 0.5$ for each model. Using our previously described procedure, we generate a total of 10,000 simulation runs for each value of n. Our experimental setup results in $\binom{6}{2} = 15$ pairwise comparisons ($m = 15$), and as there are no underlying differences between the simulated classifiers, we have 15 true null hypotheses ($m_0 = 15$). As such, if the conditions for the BH procedure are satisfied, we expect the FDR to be less than $\text{FDR}_{\max} = \frac{15}{15}q = q$. The results are shown in Figures 1 and 2, where we display the estimated FDR rates for the BH and BY procedures, for different combinations of test set sizes and values of q. Using both McNemar's test and DeLong's test, the BH procedure appears to control the FDR by keeping it below the corresponding $\text{FDR}_{\max}$ value, shown by the dashed line, in all cases—that is, for all combinations of test set sizes and q. In comparison, the BY procedure is much more conservative, with each estimated FDR value far below the $\text{FDR}_{\max}$ line.

For our second set of simulations, we use the values of σ that appear in Table 1 to generate six different models. As there are three models with the same value of $\sigma = 0.1$, we have $\binom{3}{2} = 3$ true null hypotheses ($m_0 = 3$) out of 15 total comparisons ($m = 15$). Thus, under the appropriate conditions the BH procedure should keep the FDR at or below $\text{FDR}_{\max} = \frac{3}{15}q = \frac{1}{5}q$. The results are given in Figures 3 and 4, where we can see that the estimated FDR values using the BH procedure are at or below the value of $\text{FDR}_{\max}$, given by the dashed line, in all cases—that is, for all combinations of test set sizes and q. As before, the estimated FDR values from the BY procedure are very low, with each value again appearing far below the corresponding $\text{FDR}_{\max}$ line.

These results are seemingly consistent with previous works analyzing the performance of the BH procedure with pairwise comparisons [21, 38]. The findings from several of these studies are summarized in [39], where the author states that in "all the studies, for all configurations of true and false hypotheses simulated, for balanced and for non-balanced designs, normal and non-normal distributions, the BH procedure controlled the FDR." Thus, combining these previous results with our experiments from this section, there appears to be good evidence that the BH procedure properly controls the FDR in the case of pairwise comparisons of classification models. We return to this subject in the discussion.

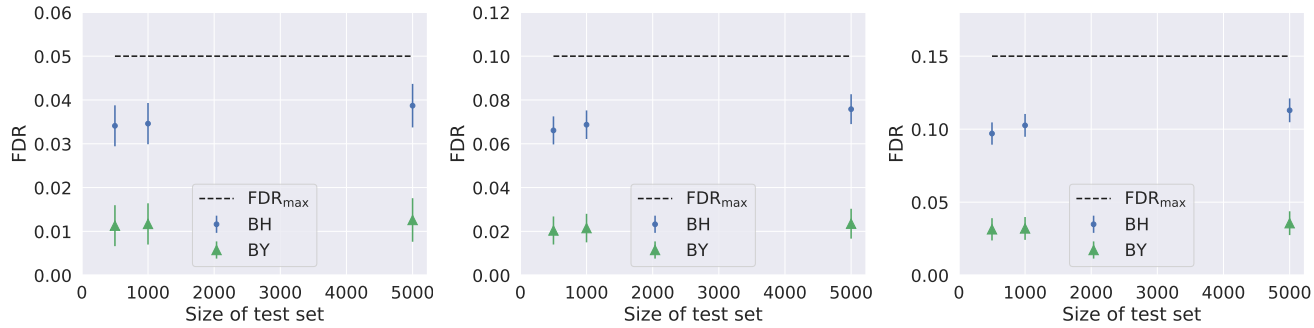
5. TRANSITIONS IN SEQUENTIAL DATA

In our second scenario we look at data that are sequential in nature, as examples of such data appear in many areas of educational research. One particular focus with sequential data is the analysis of transitions between different states—or events—in these sequences. Researchers are often interested in understanding if transitions between certain pairs of states are significant, either because they happen often or because they rarely appear. Typically in such cases, many pairs of states are evaluated with statistical tests, thus necessitating a correction for multiple comparisons. For example, past studies have analyzed logs of student actions in learning systems, in an attempt to understand the differences between productive and unproductive transitions between activities within these systems [5, 6]. Another example is affect dynamics research, which studies sequences of student affective states, with the goal of understanding how students transition between these different states. Previous works in this area have used the BH procedure to control the FDR [18, 29], and as such the goal of our next analysis is to investigate the appropriateness of using this procedure when analyzing state transitions.

5.1 Experimental Setup

Our numerical experiments for sequential data evaluate the BH and BY procedures on simulated sequences of states. Each of these sequences could represent, for example, a student's affective states while working in a learning system. The states are randomly sampled according to the probability distribution given in Table 2; each entry in the table gives the probability of sampling the next state (column) based on the value of the previous state (row). For example, 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 $0.2 - \gamma$ of being the next state, and so on.

For our simulations, we use two different values for 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

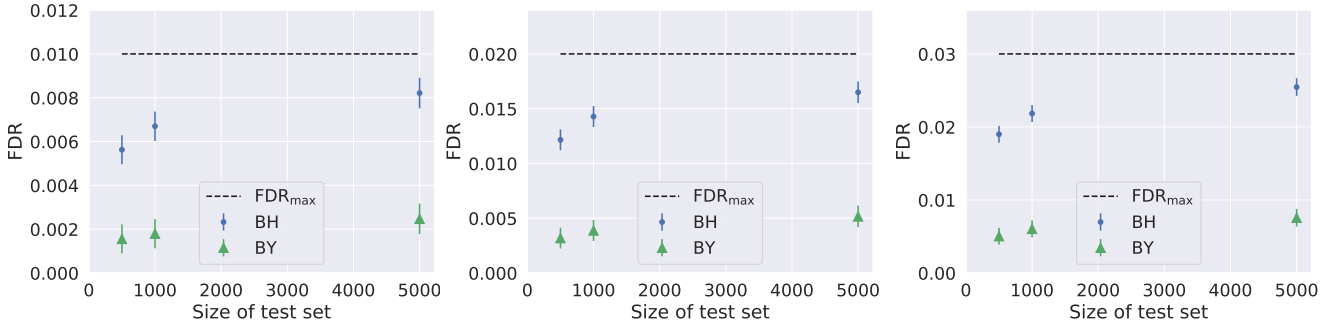


(a) $q = 0.05$ (b) $q = 0.1$ (c) $q = 0.15$

Figure 1: Comparison of the estimated FDR for the BH and BY procedures, using McNemar's test and six classifiers with the same value of $\sigma = 0.5$. Vertical lines represent the 99% confidence interval for each estimated FDR value.

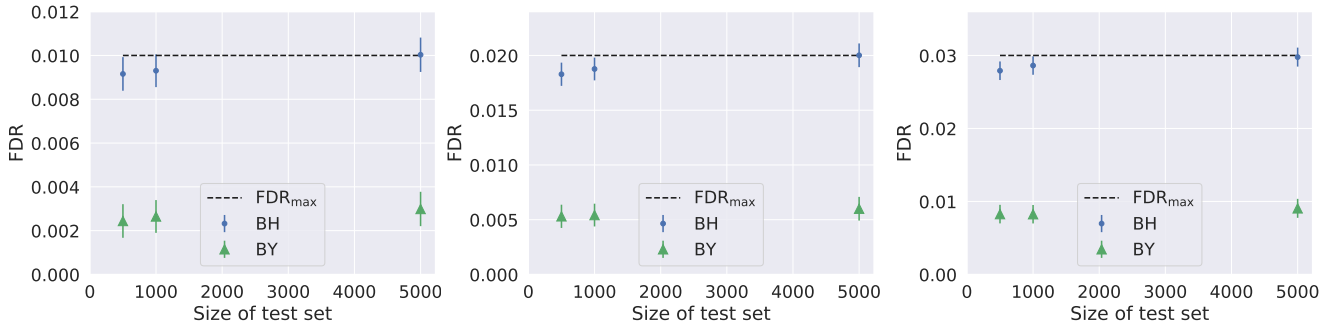
(a) $q = 0.05$ (b) $q = 0.1$ (c) $q = 0.15$

Figure 2: Comparison of the estimated FDR for the BH and BY procedures, using DeLong's test and six classifiers with the same value of $\sigma = 0.5$. Vertical lines represent the 99% confidence interval for each estimated FDR value.



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

Figure 3: Comparison of the estimated FDR for the BH and BY procedures, using McNemar's test and the σ values in Table 1. Vertical lines represent the 99% confidence interval for each estimated FDR value.



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

Figure 4: Comparison of the estimated FDR for the BH and BY procedures, using DeLong's test and the σ values in Table 1. Vertical lines represent the 99% confidence interval for each estimated FDR value.

Table 2: 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).

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

Table 3: Marginal model coefficient p-values from one simulation run using $\gamma = 0.05$. With a threshold of $q = 0.05$, both the BH and BY procedures give the same statistical significance results for this example; namely, only the four transition pairs with sample probabilities modified by γ are statistically significant.

prev \ next	A	B	C	D	E
A	0.252	0.000	0.335	0.000	0.703
B	0.496	0.365	0.327	0.864	0.252
C	0.035	0.000	0.527	0.000	0.569
D	0.260	0.652	0.080	0.980	0.889
E	0.581	0.099	0.800	0.869	0.179

choices, and then all subsequent states are sampled according to the probability distribution in Table 2. For each set of n sequences we evaluate our statistical tests (described in Sections 5.2 and 5.3) and then compute the resulting value for $\mathbf{Q}$; 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: 50, 100, and 200.

The L statistic, originally introduced in [12], is intended to be used as a measure of the significance of different pairs of transitions, and it has been widely applied in the study of affect dynamics [11, 12, 18]. Given two states A and B, it measures the likelihood of transitions from A to B while taking into account the overall frequency at which B occurs. However, several recent works have revealed issues with the use of the L statistic for the analysis of state transitions [7, 18, 19]. Thus, for our simulations we use two newer methods that have been developed in response to the problems with the L statistic. First, in Section 5.2 we look at the performance of the BH procedure when used in combination with the marginal model approach outlined in [25]. Then, in Section 5.3 we evaluate the BH procedure when it is used with the modified version of the L statistic from [24].

5.2 Marginal Model

To estimate the influence that starting in state A has on the probability of making a transition to B, in this section we use the marginal model regression procedure from [25]. In this approach, the regression model has a binary response variable, where the value of this variable is one if the next state is equal to B, and it is zero otherwise. Based on the binary response variable, we use the logit as our link function. Our predictor—or, independent—variable is also binary, with a value of one if the previous state is equal to A and zero otherwise. We can summarize this procedure 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 S represent the standard logistic function, the regression equation then has the form

$$P(y_{it} = 1 \,|\, x_{it}) = S(\beta_0 + \beta_1 x_{it}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{it})}}. \quad (3)$$

When $x_{it} = 1$ the regression model returns an estimate for $P(B \,|\, A)$, the probability of a transition to B, given that the starting state is A. Then, when $x_{it} = 0$ it returns an estimate for $P(B \,|\, \overline{A})$, the probability of a transition to B, given that the starting state is not A. Thus, to measure the importance of starting in state A, we focus on testing if the value of β_1 is significantly different from zero. This is done using a two-tailed z-test on the value of β_1 for each individual fit of the regression model.

Finally, as the sequential data used in these analyses typically take the form of repeated measurements on a student,

the result is a set of dependent—or correlated—data. To account for this dependence, as outlined in [25] we use a marginal model, based on generalized estimating equations (GEE) [17, 23], to estimate the logistic regression coefficients; in particular, we use the GEE implementation from the `statsmodels` Python library.

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 $\text{FDR}_{\max} = \frac{25}{25}q = q$ if the BH conditions are satisfied. Then, for all values of $\gamma > 0$ we would expect the FDR to be less than $\text{FDR}_{\max} = \frac{21}{25}q$, assuming the BH conditions are satisfied, as $m_0 = 21$ of the tests are true null hypotheses.

The first set of results, using a value of $\gamma = 0$, is shown in Figure 5. Here, we can see that in all cases the estimated FDR values from the BH procedure are above the theoretical upper bound of $\text{FDR}_{\max}$, shown by the dashed line. The gap is particularly notable with smaller numbers of sequences. On the other hand, the BY procedure offers much more stringent control of the FDR, with all of the estimated values appearing below the $\text{FDR}_{\max}$ line. Figure 6 then shows the results from using a value of $\gamma = 0.05$. Overall, the picture appears similar to the $\gamma = 0$ case, with the estimated FDR values from the BH procedure always appearing above the $\text{FDR}_{\max}$ line, and with the difference again being more pronounced with smaller numbers of sequences.

5.3 Removing Self-Transitions

Our final set of experiments investigates a specific situation in sequential data analysis that occurs 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 [18]. However, this procedure has been shown to overestimate the significance of transitions when used with the L statistic [19]. Thus, for this analysis we instead use a modified version of the L statistic, named L^* [24].

Definition 1. *Let A and B be two states, and let*

$$T_{\overline{A}} = \{\textit{transitions where the next state is not } A\}. \tag{4}$$

Then, we define

$$L^*(A \to B) := \frac{P(B \mid A, T_{\overline{A}}) - P(B \mid T_{\overline{A}})}{1 - P(B \mid T_{\overline{A}})}, \tag{5}$$

where $P(B \mid A, T_{\overline{A}})$ is the probability of a transition to B in $T_{\overline{A}}$, given that the starting state is A, while $P(B \mid T_{\overline{A}})$ is the overall probability of a transition to B in $T_{\overline{A}}$.

The base rate of the state B, given by $P(B \mid T_{\overline{A}})$ in (5), can be computed either individually for each sequence, or averaged over the entire set of sequences. For the computations in the remainder of this work, we compute these rates individually per sequence.

Our analysis using L^* applies the statistic to the sequences from our experiments in Section 5.2. Specifically, we take each sequence and, for each pair of transition states, compute (5). To test for statistical significance, we follow the procedure outlined in [24] and apply a two-tailed t-test to the L^* values. The results for the $\gamma = 0$ and $\gamma = 0.05$ sequences are shown in Figures 7 and 8, respectively. While perhaps not quite as prominent as with the marginal model procedure, there are several examples where the estimated FDR values from the BH procedure are clearly above the $\text{FDR}_{\max}$ line. As with the marginal model procedure, the worst cases occur with the smallest number of sequences.

5.4 Dependence of the Statistical Tests

The experiments in this section provide evidence that, when used in combination with either the marginal model procedure or L^*, the BH procedure does not always control the FDR at the desired level; in turn, this may indicate that the conditions for applying the BH procedure are not satisfied. In the remainder of this section, we outline two arguments that show the assumption of independence is violated between the statistical tests used in these analyses. Note that these are not rigorous mathematical proofs; rather, our goal here is to simply give some intuition into the relationships between the statistical tests.

Consider a set of sequential data consisting of possible states A, B, C, D, and E. For states A and B, let $\beta_{A,B}$ represent the value of β_1 in (3) for transitions of the form $A \to B$. Suppose that the following inequalities hold.

$$\begin{aligned} \beta_{A,A} &> 0 & \beta_{A,B} &> 0 \\ \beta_{A,C} &> 0 & \beta_{A,D} &> 0 \end{aligned} \tag{6}$$

Consider, for example, $\beta_{A,B}$. The corresponding marginal model estimates the probability of a transition to B, depending on whether or not the starting state is A—these estimates correspond to $P(B \mid A)$ and $P(B \mid \overline{A})$, respectively. The inequalities in (6) can then be interpreted as follows.

$$\begin{aligned} P(A \mid A) &> P(A \mid \overline{A}) & P(B \mid A) &> P(B \mid \overline{A}) \\ P(C \mid A) &> P(C \mid \overline{A}) & P(D \mid A) &> P(D \mid \overline{A}) \end{aligned} \tag{7}$$

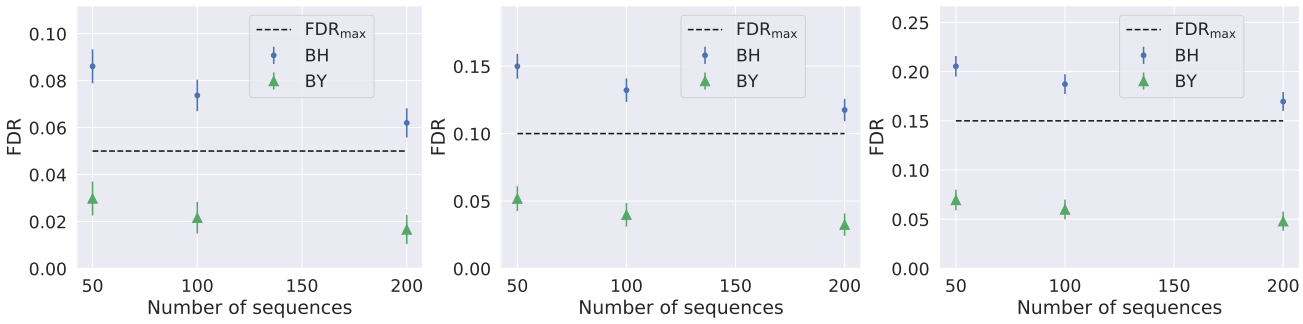
Next, consider the following two equalities.

$$\begin{aligned} P(E \mid A) &= 1 - P(A \mid A) - P(B \mid A) - P(C \mid A) - P(D \mid A) \\ P(E \mid \overline{A}) &= 1 - P(A \mid \overline{A}) - P(B \mid \overline{A}) - P(C \mid \overline{A}) - P(D \mid \overline{A}) \end{aligned} \tag{8}$$

Combining (7) and (8), it follows that $P(E \mid A) < P(E \mid \overline{A})$, or, equivalently, that $\beta_{A,E} < 0$. What this argument illustrates is that it's not possible—or, at least, it's highly unlikely—for $\beta_{A,E}$ to be positive when the other four coefficients are positive, which means that the corresponding statistical tests are not completely independent of each other.

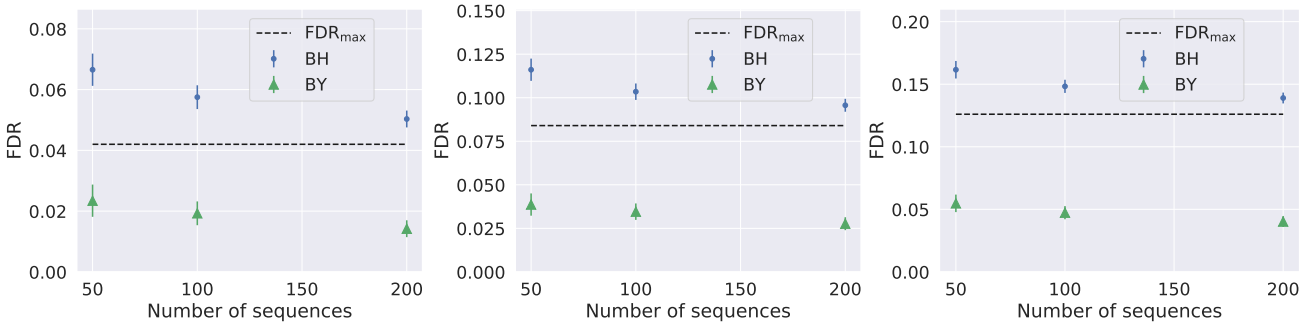
Next, suppose we are in the situation of removing self-transitions and applying L^*; thus, in what follows assume we are interested in transitions from A to B and that, following (4) in Definition 1, all transitions to A have been removed from our sequence. Suppose the following inequalities hold.

$$\begin{aligned} P(B \mid A) &> P(B) \\ P(C \mid A) &> P(C) \\ P(D \mid A) &> P(D) \end{aligned} \tag{9}$$



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

Figure 5: Comparison of the estimated FDR for the BH and BY procedures, using a value of $\gamma = 0$ and the marginal model method. Vertical lines represent the 99% confidence interval for each estimated FDR value.



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

Figure 6: Comparison of the estimated FDR for the BH and BY procedures, using a value of $\gamma = 0.05$ and the marginal model method. Vertical lines represent the 99% confidence interval for each estimated FDR value.

Consider the equalities

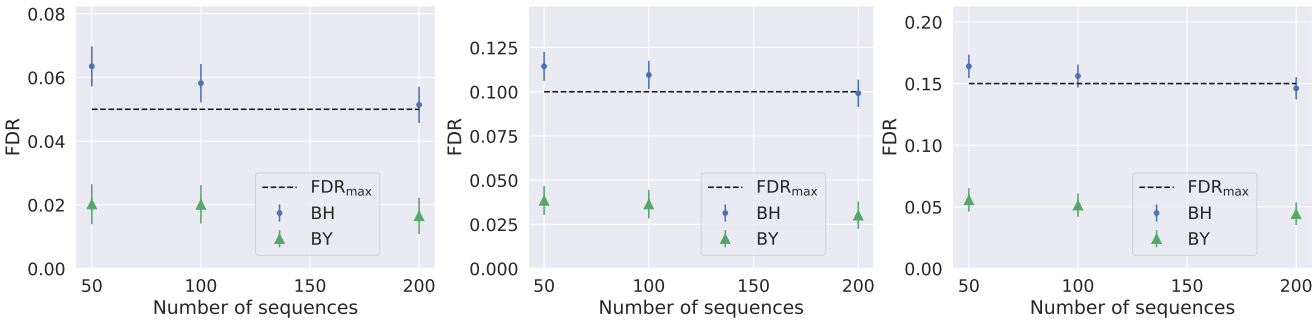
$$\begin{aligned} P(E \mid A) &= 1 - P(B \mid A) - P(C \mid A) - P(D \mid A) \\ P(E) &= 1 - P(B) - P(C) - P(D), \end{aligned} \tag{10}$$

where we're using the fact that, as we are removing transitions to A, $P(A \mid A) = 0$ and $P(A) = 0$. Combining (9) and (10), it follows that $P(E \mid A) < P(E)$. Thus, it's not possible for all four of the conditional probabilities to be larger than the base probabilities—in turn, this means that at least one of the L^* values must be negative. As such, it follows that the corresponding statistical tests are not completely independent of each other.

6. DISCUSSION

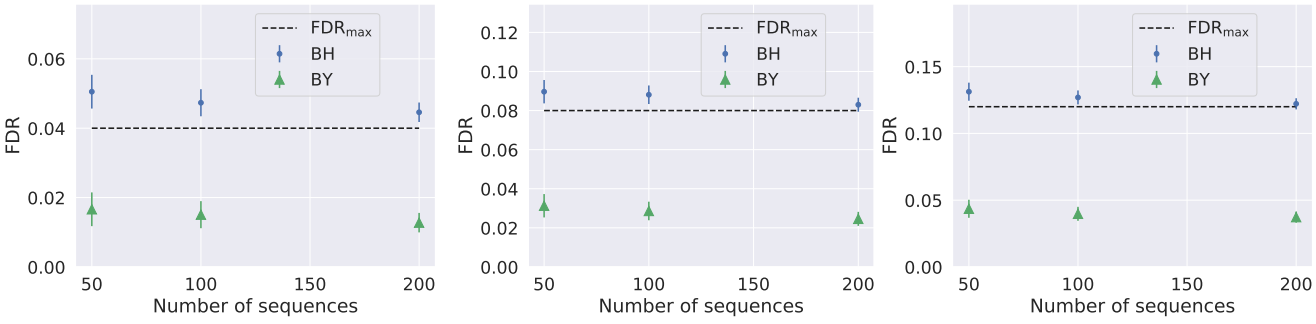
In this paper, we investigated the validity of methods used to adjust for false discoveries when performing multiple comparisons. In two scenarios relevant to EDM research, we evaluated the performance of the commonly used BH procedure in relation to an alternate method—the BY procedure—that is more general and is valid to use when the assumptions of the BH procedure cannot be met. Our first set of experiments looked at the performance of these procedures when used with pairwise comparisons of classification models on a fixed set of test data. In all our experiments, using both accuracy and AUROC as our performance metrics, the BH procedure controlled the FDR at the expected level. These results are consistent with previous studies investigating pairwise comparisons, where in all cases the BH procedure properly controlled the FDR [21, 38, 39]. Combining these previous results with the experiments in this study, our current view is that the usage of the BH procedures appears justified in this scenario—that is, one can reasonably expect the BH procedure to properly control the FDR when performing pairwise comparisons of classifiers on a fixed set of test data.

Contrast this with our investigation on sequential data, where we observed that the BH procedure, when combined with either the marginal model procedure or L^*, did not control the FDR at the expected level—this happened with various experimental conditions and for various threshold values q. The results could be an indication that the theoretical conditions for applying the BH procedure might not be satisfied in these situations. Combined with the fact that various issues involving the analysis of state transitions have recently come to light [7, 18, 19, 24, 25], we believe that using the more conservative BY procedure is justified, particularly when the analysis involves a small number of sequences. To compensate for the fact that it is more conservative, when applying the BY procedure we suggest the use of a larger value of q, such as 0.1.



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

Figure 7: Comparison of the estimated FDR for the BH and BY procedures, using a value of $\gamma = 0$ and the L^* statistic. Vertical lines represent the 99% confidence interval for each estimated FDR value.



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

Figure 8: Comparison of the estimated FDR for the BH and BY procedures, using a value of $\gamma = 0.05$ and the L^* statistic. Vertical lines represent the 99% confidence interval for each estimated FDR value.

More generally, it's worth noting that there are many examples where the BH procedure performs well without any theoretical guarantees [14, 22]. Thus, for situations in which the BH procedure has not been theoretically or empirically vetted, we offer a couple of suggestions. First, whenever possible, conducting a simulation study may be helpful; as seen in this work, the results could give evidence for or against the usage of the BH procedure. Failing that, and if there is good reason to doubt the validity of using the BH procedure, we suggest that the BY procedure be considered as a possible alternative. In these cases, a higher value for q may be justified in order to compensate for the more restrictive nature of the BY procedure, and this decision could be made based on the context of the study. For instance, in studies that are exploratory in nature or have small sample sizes, the loss of statistical power might be a larger concern; thus, the BY procedure using a threshold of 0.1 or larger may be appropriate. Whereas, in an experimental study looking for conclusive evidence, it may be preferable to use the BY procedure with a smaller value of q.

In regards to future work in this area, it would be of interest to more completely understand why the BH procedure fails to properly control the FDR in our simulations with sequential data. While we presented an argument in Section 5.4 that showed the statistical tests are not independent, it's an open question whether this argument can be extended to rigorously show that the assumptions of the BH procedure are violated—we are currently looking at this in more detail. Furthermore, it's possible that other elements may also be at play. For example, as discussed previously there are known issues with several existing methods commonly used to evaluate state transitions. While the methods we used in this study were originally developed in response to these problems [24, 25], it's possible that these existing issues, or perhaps even new ones, are a factor; thus, further adjustments to the marginal model and L^* methods could lead to improved control of the FDR with the BH procedure.

There exist other directions for future work that we are currently exploring. First, as the literature on multiple comparisons and controlling the FDR is actively growing, many methods have been developed over the years. Thus, while the BH and BY procedures are arguably the most notable of the FDR controlling procedures, it would be worthwhile to evaluate some of the newer alternatives, especially for the analysis of state transitions. Second, our analyses in this work focused exclusively on false discoveries (Type I errors) and did not consider false negatives (Type II errors). As such, in future work we aim to explicitly examine the interaction between these two types of errors with respect to the BH and BY procedures and EDM research.

7. REFERENCES

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

[2] Y. Benjamini. Selective inference: The silent killer of replicability. *Harvard Data Science Review*, 2(4), 12 2020. https://hdsr.mitpress.mit.edu/pub/l39rpgyc.

[3] Y. Benjamini and Y. Hochberg. 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, 1995.

[4] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, pages 1165–1188, 2001.

[5] G. Biswas, H. Jeong, J. Kinnebrew, B. Sulcer, and R. D. Roscoe. Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Res. Pract. Technol. Enhanc. Learn.*, 5:123–152, 2010.

[6] N. Bosch and S. D'Mello. The affective experience of novice computer programmers. *International Journal of Artificial Intelligence in Education*, 27(1):181–206, 2017.

[7] N. Bosch and L. Paquette. What's next? Edge cases in measuring transitions between sequential states. 2020. Submitted for publication.

[8] C. Cody, M. Maniktala, D. Warren, M. Chi, and T. Barnes. Does autonomy help Help? The impact of unsolicited hints and choice on help avoidance and learning. In *Proceedings of the 13th International Conference on Educational Data Mining*, pages 591–595, 2020.

[9] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, pages 837–845, 1988.

[10] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

[11] S. D'Mello and A. Graesser. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157, 2012.

[12] S. D'Mello, R. S. Taylor, and A. Graesser. Monitoring affective trajectories during complex learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society, 29 (29)*, pages 203–208, 2007.

[13] A. L. Edwards. Note on the correction for continuity in testing the significance of the difference between correlated proportions. *Psychometrika*, 13(3):185–187, 1948.

[14] A. Farcomeni. More powerful control of the false discovery rate under dependence. *Statistical Methods and Applications*, 15(1):43–73, 2006.

[15] W. Fithian and L. Lei. Conditional calibration for false discovery rate control under dependence. *arXiv preprint arXiv:2007.10438*, 2020.

[16] J. J. Goeman and A. Solari. Multiple hypothesis testing in genomics. *Statistics in medicine*, 33(11):1946–1978, 2014.

[17] P. J. Heagerty and S. L. Zeger. Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statistical Science*, 15(1):1–26, 2000.

[18] S. Karumbaiah, J. Andres, A. F. Botelho, R. S. Baker, and J. S. Ocumpaugh. The implications of a subtle difference in the calculation of affect dynamics. In *Proceedings of the 26th International Conference on Computers in Education*, pages 29–38, 2018.

[19] S. Karumbaiah, R. S. Baker, and J. Ocumpaugh. The case of self-transitions in affective dynamics. In *Artificial Intelligence in Education-20th International Conference, AIED 2019*, pages 172–181, 2019.

[20] S. Karumbaiah, J. Ocumpaugh, and R. S. Baker. The influence of school demographics on the relationship between students help-seeking behavior and performance and motivational measures. In *Proceedings of the 12th International Conference on Educational Data Mining*, pages 99–108, 2019.

[21] H. Keselman, R. Cribbie, and B. Holland. The pairwise multiple comparison multiplicity problem: An alternative approach to familywise and comparison wise Type I error control. *Psychological Methods*, 4(1):58, 1999.

[22] K. I. Kim and M. A. van de Wiel. Effects of dependence in high-dimensional multiple testing problems. *BMC bioinformatics*, 9(1):1–12, 2008.

[23] K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[24] J. Matayoshi and S. Karumbaiah. Adjusting the L statistic when self-transitions are excluded in affect dynamics. *Journal of Educational Data Mining*, 12(4):1–23, Dec. 2020.

[25] J. Matayoshi and S. Karumbaiah. Using marginal models to adjust for statistical bias in the analysis of state transitions. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, pages 449– 455, 2021.

[26] J. McDonald. *Handbook of Biological Statistics (3rd ed.)*. Sparky House Publishing, 2014.

[27] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[28] G. Nolte, O. Bai, L. Wheaton, Z. Mari, S. Vorbach, and M. Hallett. Identifying true brain interaction from eeg data using the imaginary part of coherency. *Clinical neurophysiology*, 115(10):2292–2307, 2004.

[29] J. Ocumpaugh, J. M. Andres, R. Baker, J. DeFalco, L. Paquette, J. Rowe, B. Mott, J. Lester, V. Georgoulas, K. Brawner, et al. Affect dynamics in military trainees using vMedic: From engaged concentration to boredom to confusion. In *International Conference on Artificial Intelligence in Education*, pages 238–249. Springer, 2017.

[30] S. J. Pocock, M. D. Hughes, and R. J. Lee. Statistical problems in the reporting of clinical trials. *New England Journal of Medicine*, 317(7):426–432, 1987.

[31] A. Reiner-Benaim. FDR control by the BH procedure for two-sided correlated tests with implications to gene expression data analysis. *Biometrical Journal*, 49(1):107–126, 2007.

[32] A. Reiner-Benaim, D. Yekutieli, N. E. Letwin, G. I. Elmer, N. H. Lee, N. Kafkafi, and Y. Benjamini. Associating quantitative behavioral traits with gene expression in the brain: Searching for diamonds in the hay. *Bioinformatics*, 23(17):2239–2246, 2007.

[33] S. Seabold and J. Perktold. Statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*, 2010.

[34] G. D. Smith and S. Ebrahim. Data dredging, bias, or confounding: They can all get you into the BMJ and the Friday papers, 2002.

[35] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[36] X. Sun and W. Xu. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393, 2014.

[37] R. Venant and M. d'Aquin. Towards the prediction of semantic complexity based on concept graphs. In *Proceedings of the 12th International Conference on Educational Data Mining*, pages 188–197, 2019.

[38] V. S. Williams, L. V. Jones, and J. W. Tukey. 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, 1999.

[39] D. Yekutieli. False discovery rate control for non-positively regression dependent test statistics. *Journal of Statistical Planning and Inference*, 138(2):405–415, 2008.

[40] S. S. Young and A. Karr. Deming, data and observational studies: A process out of control and needing fixing. *Significance*, 8(3):116–120, 2011.

8. EDM REVIEW REFERENCES

[41] H. Anderson, A. Boodhwani, and R. Baker. Assessing the fairness of graduation predictions. In *Proceedings of the 12th International Conference on Educational Data Mining*, 2019.

[42] J. Andrews-Todd, C. Forsyth, J. Steinberg, and A. Rupp. Identifying profiles of collaborative problem solvers in an online electronics environment. In *Proceedings of the 11th International Conference on Educational Data Mining*, 2018.

[43] A. Bauer, J. Flatten, and Z. Popovic. Analysis of problem-solving behavior in open-ended scientific-discovery game challenges. In *Proceedings of the 10th International Conference on Educational Data Mining*, 2017.

[44] N. Bosch, W. Crues, and N. Shaik. Diverse learners, diverse motivations: Exploring the sentiment of learning objectives. In *Proceedings of the 11th International Conference on Educational Data Mining*, 2018.

[45] A. F. Botelho, R. Baker, J. Ocumpaugh, and N. Heffernan. Studying affect dynamics and chronometry using sensor-free detectors. In *Proceedings of the 11th International Conference on Educational Data Mining*, 2018.

[46] C. Cody, M. Maniktala, D. Warren, M. Chi, and T. Barnes. Does autonomy help Help? The impact of unsolicited hints and choice on help avoidance and learning. In *Proceedings of the 13th International Conference on Educational Data Mining*, 2020.

[47] M. Dong, R. Yu, and Z. Pardos. Design and deployment of a better university course search: Inferring latent keywords from enrollments. In *Proceedings of the 12th International Conference on Educational Data Mining*, 2019.

[48] M. Eagle, A. Corbett, J. Stamper, and B. McLaren. Predicting individualized learner models across tutor lessons. In *Proceedings of the 11th International Conference on Educational Data Mining*, 2018.

[49] E. Farhana, T. Rutherford, and C. Lynch. Investigating relations between self-regulated reading behaviors and science question difficulty. In *Proceedings of the 13th International Conference on Educational Data Mining*, 2020.

[50] S. C. Fonseca, F. D. Pereira, E. Oliveira, D. Fernandes, L. S. D. Carvalho, and A. Cristea. Automatic subject-based contextualisation of programming assignment lists. In *Proceedings of the 13th International Conference on Educational Data Mining*, 2020.

[51] C. Forsyth, J. Andrews-Todd, and J. Steinberg. Are you really a team player? profiling of collaborative problem solvers in an online environment. In *Proceedings of the 13th International Conference on Educational Data Mining*, 2020.

[52] P. S. Inventado, P. Scupelli, E. V. Inwegen, K. S. Ostrow, N. Heffernan, J. Ocumpaugh, R. Baker, S. Slater, and M. Almeda. Hint availability slows completion times in summer work. In *Proceedings of the Ninth International Conference on Educational Data Mining*, 2016.

[53] S. Klingler, R. Wampfler, T. Käser, B. Solenthaler, and M. Gross. Efficient feature embeddings for student classification with variational auto-encoders. In *Proceedings of the 10th International Conference on Educational Data Mining*, 2017.

[54] Z. Liu, R. Brown, C. Lynch, T. Barnes, R. Baker, Y. Bergner, and D. McNamara. MOOC learner behaviors by country and culture; an exploratory analysis. In *Proceedings of the Ninth International Conference on Educational Data Mining*, 2016.

[55] Z. Liu, C. Cody, T. Barnes, C. Lynch, and T. Rutherford. The antecedents of and associations with elective replay in an educational game: Is replay worth it? In *Proceedings of the 10th International Conference on Educational Data Mining*, 2017.

[56] I. Pytlarz, S. Pu, M. Patel, and R. Prabhu. What can we learn from college students' network transactions? Constructing useful features for student success prediction. In *Proceedings of the 11th International Conference on Educational Data Mining*, 2018.

[57] S. Slater, J. Ocumpaugh, R. Baker, P. Scupelli, P. S. Inventado, and N. Heffernan. Semantic features of math problems: Relationships to student learning and engagement. In *Proceedings of the Ninth International Conference on Educational Data Mining*, 2016.

[58] K. Thaker, Y. Huang, P. Brusilovsky, and H. Da-qing. Dynamic knowledge modeling with heterogeneous activities for adaptive textbooks. In *Proceedings of the*

11th International Conference on Educational Data Mining, 2018.

[59] A. Vail, J. Wiggins, J. F. Grafsgaard, K. Boyer, E. Wiebe, and J. Lester. The affective impact of tutor questions: Predicting frustration and engagement. In *Proceedings of the Ninth International Conference on Educational Data Mining*, 2016.

[60] O. Vainas, Y. B. David, R. Gilad-Bachrach, M. Ronen, O. Bar-Ilan, R. Shillo, G. Lukin, and D. Sitton. Staying in the zone: Sequencing content in classrooms based on the zone of proximal development. In *Proceedings of the 12th International Conference on Educational Data Mining*, 2019.

[61] R. Venant and M. d'Aquin. Towards the prediction of semantic complexity based on concept graphs. In *Proceedings of the 12th International Conference on Educational Data Mining*, 2019.

[62] R. Zhi, S. Marwan, Y. Dong, N. Lytle, T. W. Price, and T. Barnes. Toward data-driven example feedback for novice programming. In *Proceedings of the 12th International Conference on Educational Data Mining*, 2019.