A New Framework for Online Testing of Heterogeneous Treatment Effect

Miao Yu, Wenbin Lu, Rui Song
Department of Statistics
North Carolina State University
{myu12, wlu4, rsong}@ncsu.edu

Abstract

We propose a new framework for online testing of heterogeneous treatment effects. The proposed test, named sequential score test (SST), is able to control type I error under continuous monitoring and detect multi-dimensional heterogeneous treatment effects. We provide an online p-value calculation for SST, making it convenient for continuous monitoring, and extend our tests to online multiple testing settings by controlling the false discovery rate. We examine the empirical performance of the proposed tests and compare them with a state-of-art online test, named mSPRT using simulations and a real data. The results show that our proposed test controls type I error at any time, has higher detection power and allows quick inference on online A/B testing.

1 Introduction

Randomized controlled experiment, also known as *A/B testing*, is widely used in web facing industry to improve products and technologies in a data-driven manner (Kohavi et al. 2009). Most of A/B tests are conducted by performing a formal *null hypothesis statistical testing* (NHST) with the typical *null hypothesis* $H_0 : \beta := \mu_B - \mu_A = 0$ to determine if the difference of the metric across two variants is significant or not. The result of a NHST is summarized in a *p-value* and the case that the p-value is less than a preset *significance level* α will lead the null hypothesis to be rejected. A valid testing is able to get a high power to detect the difference if there is, while controlling the *type I error*, i.e., the probability of erroneously rejecting H_0, to be less than α.

However, the validity of NHST requires that the sample size is fixed in advance, which is often violated in practice. In A/B testing practice, a fast-paced product evolution pushes its shareholders to continuously monitor the p-values and draw conclusions prematurely. In fact, stopping experiments in an adaptive manner can favorably bias getting significant results and lead to very high false positive probabilities, well in excess of the nominal significance level (Goodson 2014; Simmons, Nelson, and Simonsohn 2011). As an extreme example in (Pekelis, Walsh, and Johari 2015), it can be shown that stopping the first time that the p-value is less than α actually has type I error probability of 1. Yet for all that, this "peeking" behavior is not without reasons. The time cost and opportunity cost for *fixed-horizon* hypothesis testing are large (Ju et al. 2019), so users want to find true effects and stop the experiments as quickly as possible. Moreover, the sample size calculation requires an estimate of the *minimum detectable effect* (MDE). Most users lack good prior knowledge of the trade-off between high detection ability and short waiting time and may want to adjust them after peeking early at results.

Another problem of A/B testing is that it assumes there is only an *average treatment effect* (ATE) in the population of experiment. But underlying this average effect may be substantial variation in how particular subgroups respond to treatments: there may be *heterogeneous treatment effects* (HTE) (Grimmer, Messing, and Westwood 2017). It might be that the population average effect of a product with a new feature is not significant, but the feature does benefit a lot among particular subgroups of users. In this case, we won't be able to detect those effects and will lose the chance of making profits by promoting new products to those target sub-populations, if only ATE is tested in A/B testing.

To address the continuously monitoring problem, *sequential testing* (ST) was first developed by Wald (Wald 1945), who introduced the *sequential probability ratio test* (SPRT). ST allows intermediate checks of significance while providing type I error control at any time; see (Lai 2001) for a survey on sequential testing. Moreover, ST could help decision makers conclude an experiment earlier with often much fewer samples than the fixed-horizon testing (Wald 1945). *Mixture sequential probability ratio test* (mSPRT) (Robbins 1970) and *maximized sequential probability test* (MaxSPRT) (Kulldorff et al. 2011) are two variants of sequential testing that generalized SPRT to a composite hypothesis. Due to the merits of mSPRT that it is a test with power 1 (Robbins and Siegmund 1974) and almost optimal (Pollak 1978) with respect to expected time to stop, it was brought to A/B testing by Johari et al. (2015; 2017). They also proposed a notion of *always valid p-value process* (sequential p-values) in the same papers and used it as a tool for converting fixed-horizon *multiple testing* procedures to a sequential version. Later, Malek et al. (2017) also showed that if the original multiple testing procedure has a type I error guarantee in a certain family (including false discovery rate and family-wise error rate), then the sequential conversion inherits an analogous guarantee.

However, current online testing procedures, such as mSPRT, are not suitable for testing heterogeneous treatment effects due to two aspects. First, they can not accommodate the nuisance parameters in the baseline effects. Second, they may not be able to control the type I error and may lack of power for detecting heterogeneous treatment effects. In this paper, we propose a new framework for online testing of heterogeneous treatment effects. The proposed test, named SST, is based on the ratio of asymptotic score statistic distributions, which is able to test multi-dimensional parameters. Furthermore, the asymptotic normality of the score functions guarantees an explicit form of the integral, which allows the integration for the ratio to be efficient. At last, we generalize our framework to online multiple testing, which is often the case in industrial practice.

The remainder of this paper is structured as follows. In Section 2, we introduce some preliminary knowledge about fixed-horizon testing and sequential testing. In Section 3, we present the proposed new framework for online testing of heterogeneous treatment effects. We extend SST to multiple testing settings in Section 4 and conduct experiments in Section 5 to compare our framework with the widely-used mSPRT. Finally, in Section 6, we conclude the paper and present future directions.

2 Preliminaries

2.1 Fixed-horizon testing

Fixed-horizon testing is the most widely used procedure in industry where the sample size is fixed in advance. It can be broken down into several steps (Lehmann and Romano 2006):

Step 1: Determine a desired significance level α, minimum detectable effect (MDE) and power at MDE. It means that the probability to detect the MDE is at least at the value of power, while the probability of rejecting H_0, if it is actually true, is at most α.

Step 2: Calculate/Estimate the minimum sample size n. The sample size n needs to be large enough to achieve the desired power at MDE while controlling type I error at a significance level, but too large sample size will lead to more opportunity cost of waiting for more samples. One need to trade off between these two aspects when choosing sample size.

Step 3: Collect n samples and compute the observed value of an appropriate test statistics Λ_n. The most common test statistics for two-sample tests are z-tests and t-tests, which assume that data are from a normal distribution with known or unknown variance, respectively.

Step 4: Compute a p-value p_n and reject the null hypothesis if $p_n \leq \alpha$. P-value is a random variable to denote the probability of seeing a test statistic as extreme as the observed statistics Λ_n under null hypothesis, and can be formally defined as

$$p_n = \inf\{\alpha : \Lambda_n \geq k(\alpha)\}, \tag{1}$$

where $k(\alpha)$ is a critical value depending on significance level and the distribution of Λ_n under H_0. The critical value is determined such that, under the null hypothesis H_0, the event $\Lambda_n \geq k(\alpha)$ occurs with probability no greater than α. Since the p-value was computed assuming a fixed sample size n, we refer to this as a *fixed-horizon p-value*. Small p-values suggest evidence in support of alternative hypothesis.

A *decision rule* is a pair (T, δ) representing a testing, where T is a stopping time indicating the sample size at which the test is ended, and δ is a binary indicator for rejection decision. With the definition of fixed-horizon p-value in (1), it is obvious to see that (n, δ_1) and (n, δ_2) with $\delta_1 = 1\{p_n \leq \alpha\}$ and $\delta_2 = 1\{\Lambda_n \geq k(\alpha)\}$ are two equivalent decision rules for fixed-horizon testing. That means the decision rule and p-value can be obtained from each other: find p-value from decision rule (n, δ_2) by (1), or make the decision (n, δ_1) from p-value. Hence, we can actually stop at step 3 and reject H_0 if $\Lambda_n \geq k(\alpha)$ for some predetermined significance level α. Nonetheless, the decision-making process using p-values is remarkably simple and transparent: one can choose their own significance level and make a valid decision.

2.2 Sequential testing

Sequential testing, contrast to fixed-horizon, is a procedure where the decision of terminating the process at any stage of the experiment depends on the results of the observations previously made. It has gained recent popularity in online A/B testing (Balsubramani and Ramdas 2015; Johari, Pekelis, and Walsh 2015) due to its flexibility of continuously monitoring and ending the experiment as soon as significant results are observed.

The decision rules for sequential testing is a nested family of $(T(\alpha), \delta(\alpha))$, parameterized by significance level α. It has the following two properties (Johari et al. 2017): First, the type I error is controlled, that is, $P_{H_0}(\delta(\alpha) = 1) \leq \alpha$; Second, $T(\alpha)$ is (almost surely) non-increasing in α while $\delta(\alpha)$ is (almost surely) non-decreasing in α. In other words, less stringent type I error control allows the test to stop sooner, and is more likely to lead to rejection.

Similar to fixed-horizon testing, a notion of sequential p-values was also introduced for sequential testing and named *always valid p-value process* by (Johari et al. 2017): *A sequence of fixed-horizon p-values $(p_n)_{n=1}^{\infty}$ is always valid if it satisfies the property that $\forall s \in [0, 1]$, $\mathbb{P}_{H_0}(p_T \leq s) \leq s$ for any given (possibly infinite) stopping time T.* It allows the user to trade off detection power and sample size dynamically as they see fit while still control type I error. In the same way, the always valid p-values can be derived from the decision rule for a sequential test, and vice versa. For a given sequential test $(T(\alpha), \delta(\alpha))$,

$$p_n = \inf\{\alpha : T(\alpha) \leq n, \delta(\alpha) = 1\} \tag{2}$$

defines an always valid p-value process. For any always valid p-value process $(p_n)_{n=1}^{\infty}$, a sequential test is obtained as follows:

$$T(\alpha) = \inf\{n : p_n \leq \alpha\} \quad \delta(\alpha) = 1\{T(\alpha) < \infty\}. \tag{3}$$

The *mixture sequential probability ratio test* (mSPRT) (Robbins 1970) is a well studied family of sequential tests. Its test statistic based on the first n observations Λ_n^{π} is a mixture of likelihood ratios against the null hypothesis, with the

mixture density $\pi(\cdot)$ over the space for target parameter β. The decision rule for mSPRT is as below:

$$T(\alpha) = \inf\{n : \Lambda_n^\pi \geq \alpha^{-1}\} \qquad \delta(\alpha) = 1(T(\alpha) < \infty). \tag{4}$$

It can be shown that the type I error for mSPRT is well controlled at α by a simple application of *optional stopping theorem* (Grimmett, Grimmett, and Stirzaker 2001), since the likelihood ratio under H_0 is a nonnegative martingale with initial value equal to one and so is the mixture of such likelihood ratios; see (Malek et al. 2017; Pekelis, Walsh, and Johari 2015) for a detailed proof.

Johari et al. (2017), recently, have brought mSPRT to online A/B tests where testing parameters μ_A, μ_B are assumed to be the mean of Bernoulli or normal distribution, depending on whether the data is binary or continuous. They modified the original mSPRT to make it applicable to industrial A/B tests based on some approximation techniques, and empirically showed that the new test has high detection performance with type I error control.

2.3 Heterogeneous Treatment Effect

Up to now, all the online A/B tests we have talked about are focusing only on testing the *average treatment effect* (ATE). However, treatment effects are commonly believed to be varying among individuals, and individual treatment effects may differ in magnitude and even have opposite direction. This is called *heterogeneous treatment effect* (HTE). Testing HTE could help us identify sub-populations where treatment shows better performance and allow personalized treatment as well.

To give a better insight of the difference between ATE and HTE testing, let's take the generalized linear model (GLM) for example,

$$\begin{aligned} &Y_i \overset{ind}{\sim} \text{Exponential Family}(\gamma_i, \phi), \quad i = 1, \cdots, n \\ &f_{Y_i}(y_i|\gamma_i, \phi) = \exp\left\{\frac{y_i\gamma_i - b(\gamma_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \end{aligned} \tag{5}$$

where n denotes the sample size, $a_i(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$ are known functions, γ_i is the *canonical parameter*, and ϕ is a typically known *dispersion parameter*. They are related to the mean and variance of the response through:

$$\mu_i = \mathbb{E}(Y_i) = b'(\gamma_i), \qquad Var(Y_i) = a_i(\phi) \cdot b''(\gamma_i). \tag{6}$$

A link function $g(\cdot)$ provides the relationship between the linear predictor and the mean of response:

$$g(\mu_i) = \eta_i. \tag{7}$$

where the linear predictor η_i has different forms depending on either ATE or HTE setting. There is always a well-defined *canonical link* derived from the response's density function, which is a link function such that $g(\mu_i) = \gamma_i$. For example, normal distribution has an identity function $g(\mu_i) = \mu_i$ as the canonical link, Bernoulli has a logit link $g(\mu_i) = \log\frac{\mu_i}{1-\mu_i}$ and Poisson has a log link $g(\mu_i) = \log\mu_i$.

HTE and ATE testings have different assumptions about the form of the linear predictor. ATE testing assumes that

$$\eta_i = \theta + \beta A_i, \tag{8}$$

and test $H_0 : \beta = \beta_0$, whereas HTE testing assumes that

$$\eta_i = \boldsymbol{\theta}^T\mathbf{X}_i + (\boldsymbol{\beta}^T\mathbf{X}_i)A_i, \tag{9}$$

and test $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, where $\mathbf{X}_i$ denotes the covariates vector with the first element being 1 indicating the intercept, and A_i denotes the binary treatment. Note that $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ in HTE testing are both vectors since at least one covariate is considered.

In the case of HTE testing, mSPRT does not work well for the following reasons:

1. The test statistic may not have an explicit form if a conjugate prior $\pi(\cdot)$ for likelihood ratio doesn't exist, as is often the case in HTE testing, e.g., logistic regression. As a result, the computation is inefficient to implement in a streaming environment;
2. The nuisance parameter $\boldsymbol{\theta}$ in the likelihood function is unknown. Even though it can be replaced by its estimator, the resulting test statistics is no longer a martingale and hence the type I error cannot be controlled. Johari, Pekelis, and Walsh (2015) used a sufficient statistic for nuisance parameter and applied *central limit theory* to deal with this issue in A/B tests with Bernoulli distribution. However, this technique failed to be extended to HTE setting.

Therefore, we want to develop a valid online test that can deal with heterogeneous treatment effect.

3 A New Framework of Sequential Testing

In this section, we propose a new framework of sequential testing, called *Sequential Score Test* (SST), which is able to test heterogeneous treatment effect while accounting for unknown individual effects. This framework is applicable to independent observations from an exponential family, which includes a large set of commonly used distributions.

Instead using integrated likelihood ratios as in mSPRT, we consider the integration of the ratios of asymptotic score statistic distributions under the local alternative against the null hypothesis. The proposed method can naturally handle nuisance parameters in testing HTE. In addition, the asymptotic representation of the score statistics under the local alternative and the null hypotheses (established in Lemma 3.1) can lead to a martingale structure under the null similarly as for the integrated likelihood ratio statistics, and the resulting test statistic have a closed form for integration, which facilitates the implementation of the proposed testing procedure.

3.1 Sequential Score Test

Suppose we have i.i.d. data $(Y_i, A_i, \mathbf{X}_i)$, where Y, A, $\mathbf{X}$ respectively denote response, binary treatment and $(p+1)$-dimensional covariates vector including an intercept, respectively. We assume that the distribution of Y_i conditional on $(A_i, \mathbf{X}_i)$ is an exponential family defined in (5)-(7) with η_i in the form of (9), where $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ denote the heterogeneous treatment effect and baseline effect, respectively. We want to test null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ against local alternative $H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 + \frac{\boldsymbol{\delta}}{\sqrt{n}} \quad (\boldsymbol{\delta} \neq 0)$.

To introduce the test statistic of SST, let's start with some notations. For ease of exposition, we suppose that each group has n observations. Let $\mathbf{S}^{(1)}_{n,\boldsymbol{\beta}}(\boldsymbol{\theta}, \boldsymbol{\beta}_0)$ denotes the score function of $\boldsymbol{\beta}$ for treatment group ($A = 1$) under the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$:

$$\mathbf{S}^{(1)}_{n,\boldsymbol{\beta}}(\boldsymbol{\theta}, \boldsymbol{\beta}_0) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i^{(1)}(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}^T} \cdot \frac{(Y_i^{(1)} - \mu_i^{(1)}(\boldsymbol{\beta}, \boldsymbol{\theta}))}{a_i(\phi) \cdot V_i^{(1)}(\boldsymbol{\beta}, \boldsymbol{\theta})} \right) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \tag{10}$$

and $\mathbf{S}^{(0)}_{n,\boldsymbol{\theta}}(\boldsymbol{\theta})$ denotes the score function of $\boldsymbol{\theta}$ for control group ($A = 0$):

$$\mathbf{S}^{(0)}_{n,\boldsymbol{\theta}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial \mu_i^{(0)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \cdot \frac{(Y_i^{(0)} - \mu_i^{(0)}(\boldsymbol{\theta}))}{a_i(\phi) \cdot V_i^{(0)}(\boldsymbol{\theta})}, \tag{11}$$

where $\mu_i^{(0)}(\boldsymbol{\theta}) = \mathbb{E}(Y_i | A_i = 0, \mathbf{X}_i)$, $\mu_i^{(1)}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathbb{E}(Y_i | A_i = 1, \mathbf{X}_i)$, $a_i(\phi) \cdot V_i^{(0)}(\boldsymbol{\theta}) = Var(Y_i | A_i = 0, \mathbf{X}_i)$ and $a_i(\phi) \cdot V_i^{(1)}(\boldsymbol{\beta}, \boldsymbol{\theta}) = Var(Y_i | A_i = 1, \mathbf{X}_i)$. For simplicity, let's assume $a_i(\phi) = a(\phi)$ for all i and $a(\phi)$ is known.

Consider the following estimated average score $\bar{\mathbf{S}}_n$ for treatment group (A=1) under $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$:

$$\bar{\mathbf{S}}_n \coloneqq \frac{1}{n} \mathbf{S}^{(1)}_{n,\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}_0), \tag{12}$$

where $\hat{\boldsymbol{\theta}}_n$ is the *maximum likelihood estimator* of $\boldsymbol{\theta}$ calculated based on data from the control group ($A = 0$). The idea behind SST is to consider the test statistic as a mixture of asymptotic probability ratios of $\bar{\mathbf{S}}_n$, instead of the likelihood ratios, under alternative hypothesis to that under null hypothesis. The test statistic $\tilde{\Lambda}^{\pi}_n$ is defined as below:

$$\tilde{\Lambda}^{\pi}_n = \int \frac{\boldsymbol{\psi}_{\left(\bar{\mathbf{I}}_n^{(1)}(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\beta}-\boldsymbol{\beta}_0), \frac{\mathbf{V}_n(\hat{\boldsymbol{\theta}}_n)}{n}\right)}(\bar{\mathbf{S}}_n)}{\boldsymbol{\psi}_{\left(\mathbf{0}, \frac{\mathbf{V}_n(\hat{\boldsymbol{\theta}}_n)}{n}\right)}(\bar{\mathbf{S}}_n)} \pi(\boldsymbol{\beta}) d\boldsymbol{\beta}, \tag{13}$$

where

- $\boldsymbol{\psi}_{(\boldsymbol{\mu}, \boldsymbol{\Sigma})}(\cdot)$ denotes the probability density function of multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$
- $\mathbf{V}_n(\boldsymbol{\theta}) = \bar{\mathbf{I}}_n^{(1)}(\boldsymbol{\theta}) + \bar{\mathbf{I}}_n^{(1)}(\boldsymbol{\theta}) \left[\bar{\mathbf{I}}_n^{(0)}(\boldsymbol{\theta})\right]^{-1} \bar{\mathbf{I}}_n^{(1)}(\boldsymbol{\theta})$
- $\bar{\mathbf{I}}_n^{(1)}(\boldsymbol{\theta}) = -\frac{1}{n} \frac{\partial \mathbf{S}^{(1)}_{n,\boldsymbol{\beta}}(\boldsymbol{\theta}, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\frac{\partial \mu_i^{(1)}(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \cdot \frac{\partial \mu_i^{(1)}(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}}}{a(\phi) \cdot V_i^{(1)}(\boldsymbol{\beta}, \boldsymbol{\theta})} \right] \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$
- $\bar{\mathbf{I}}_n^{(0)}(\boldsymbol{\theta}) = -\frac{1}{n} \frac{\partial \mathbf{S}^{(0)}_{n,\boldsymbol{\theta}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\frac{\partial \mu_i^{(0)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \cdot \frac{\partial \mu_i^{(0)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}{a(\phi) \cdot V_i^{(0)}(\boldsymbol{\theta})} \right]$
- $\pi(\cdot)$ is a "mixture" distribution over the parameter space denoting the distribution of true effects $\boldsymbol{\beta}$. It is assumed to be positive everywhere. For ease of computation, we often choose $\boldsymbol{\beta} \sim \mathbf{MVN}(\boldsymbol{\beta}_0, \tau^2 \mathbf{I})$, where $\mathbf{I}$ denotes $(p+1) \times (p+1)$ identity matrix and τ is chosen based on historical data

Intuitively, large value of $\tilde{\Lambda}^{\pi}_n$ represents the evidence against H_0 in favor of a mixture of alternatives $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, weighted by $\boldsymbol{\beta} \sim \pi(\cdot)$. The decision rule for SST is quite simple and is shown in (14). That is, given a significance level α, the test stops and rejects the null hypothesis at the first time that $\tilde{\Lambda}^{\pi}_n \geq \alpha^{-1}$; if no such time exists, it accepts the null hypothesis.

$$T(\alpha) = \inf\{n : \tilde{\Lambda}^{\pi}_n \geq \alpha^{-1}\} \quad \delta(\alpha) = 1(T(\alpha) < \infty). \tag{14}$$

The corresponding sequential (always valid) p-value at sample size n, by definition of (2), is the reciprocal of the maximum value of $\tilde{\Lambda}^{\pi}_n$ up to n:

$$p_n = \frac{1}{\max_{m \leq n} \tilde{\Lambda}^{\pi}_m}. \tag{15}$$

It is obvious to see that the online p-value is monotonically non-increasing in n and $p_{T(\alpha)} = \alpha$.

3.2 Validity of SST

The intuition of $\tilde{\Lambda}^{\pi}_n$ being the appropriate test statistics comes from representing the mixture of asymptotic probability ratios of $\bar{\mathbf{S}}_n$. In this section, we will give the asymptotic distribution of $\bar{\mathbf{S}}_n$ under null hypothesis and local alternative hypothesis, respectively. Meanwhile, we will offer some insights to demonstrate the approximate validity of SST, that is, the type I error is controlled at large sample size.

The following lemma provides the asymptotic distributions of $\bar{\mathbf{S}}_n$ with proof shown in the supplemental material.

Lemma 3.1 *For generalized linear model in (5)-(7)(9) and $\bar{\mathbf{S}}_n$ in (12), define the information matrix for each group as below:*

$$\mathbf{I}^{(0)}(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{(\mathbf{X}, \mathbf{Y})} \left[\bar{\mathbf{I}}_n^{(0)}(\boldsymbol{\theta})\right] = \mathbb{E}_{(\mathbf{X}, \mathbf{Y})} \left[\frac{\frac{\partial \mu_1^{(0)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \cdot \frac{\partial \mu_1^{(0)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}{a(\phi) \cdot V_1^{(0)}(\boldsymbol{\theta})} \right] \tag{16}$$

$$\mathbf{I}^{(1)}(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{(\mathbf{X}, \mathbf{Y})} \left[\bar{\mathbf{I}}_n^{(1)}(\boldsymbol{\theta})\right] = \mathbb{E}_{(\mathbf{X}, \mathbf{Y})} \left[\frac{\frac{\partial \mu_1^{(1)}(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \cdot \frac{\partial \mu_1^{(1)}(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}}}{a(\phi) \cdot V_1^{(1)}(\boldsymbol{\beta}, \boldsymbol{\theta})} \right] \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \tag{17}$$

Then, under null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$,

$$\sqrt{n} \bar{\mathbf{S}}_n \xrightarrow[H_0]{d} \mathbf{MVN}_{p+1}(\mathbf{0}, \mathbf{V}(\boldsymbol{\theta}_0)) \tag{18}$$

whereas under local alternative $H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 + \frac{\boldsymbol{\delta}}{\sqrt{n}}$,

$$\sqrt{n} \left(\bar{\mathbf{S}}_n - \mathbf{I}^{(1)}(\boldsymbol{\theta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right) \xrightarrow[H_1]{d} \mathbf{MVN}_{p+1}(\mathbf{0}, \mathbf{V}(\boldsymbol{\theta}_0)) \tag{19}$$

where $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{I}^{(1)}(\boldsymbol{\theta}) + \mathbf{I}^{(1)}(\boldsymbol{\theta}) \left[\mathbf{I}^{(0)}(\boldsymbol{\theta})\right]^{-1} \mathbf{I}^{(1)}(\boldsymbol{\theta})$, and $\boldsymbol{\theta}_0$ is the true value of the nuisance parameter.

By Lemma 3.1, the asymptotic probability ratio of $\bar{\mathbf{S}}_n$ under local alternative $H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 + \frac{\boldsymbol{\delta}}{\sqrt{n}}$ against under null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ can be represented as:

$$\lambda_n = \frac{\boldsymbol{\psi}_{\left(\mathbf{I}^{(1)}(\boldsymbol{\theta}_0)(\boldsymbol{\beta}-\boldsymbol{\beta}_0), \frac{\mathbf{V}(\boldsymbol{\theta}_0)}{n}\right)}(\bar{\mathbf{S}}_n)}{\boldsymbol{\psi}_{\left(\mathbf{0}, \frac{\mathbf{V}(\boldsymbol{\theta}_0)}{n}\right)}(\bar{\mathbf{S}}_n)} \tag{20}$$

Different from likelihood ratio, λ_n is not an exact martingale, but we can show that the approximate martingale property does hold when the sample size n is large enough. See the following remark for mathematical expression; the proof can be found in the supplemental material.

Remark 3.1 *For generalized linear model in (5)-(7)(9) and λ_n defined by (20), let $\mathcal{F}_n$ denote the filtration that contains historical information as below:*

$$\mathcal{F}_n = \{(\mathbf{X}_i^{(j)}, Y_i^{(j)}), i = 1, \cdots, n;\ j = 0, 1\} \tag{21}$$

Then, under the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, $\mathbb{E}\left[\lambda_{n+1}|\mathcal{F}_n\right]$ is approximately equal to $\lambda_n \cdot \exp\{o_p(1)\}$.

For practical purpose, we usually replace λ_n with its following empirical version $\tilde{\lambda}_n$:

$$\tilde{\lambda}_n = \frac{\boldsymbol{\psi}_{\left(\bar{\mathbf{I}}_n^{(1)}(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\beta}-\boldsymbol{\beta}_0),\, \frac{\mathbf{V}_n(\hat{\boldsymbol{\theta}}_n)}{n}\right)}(\bar{\mathbf{S}}_n)}{\boldsymbol{\psi}_{\left(\mathbf{0},\, \frac{\mathbf{V}_n(\hat{\boldsymbol{\theta}}_n)}{n}\right)}(\bar{\mathbf{S}}_n)} \tag{22}$$

which is exactly the main term in the definition (13) of $\tilde{\Lambda}_n^{\pi}$. The empirical ratio $\tilde{\lambda}_n$ shares the same martingale property as λ_n when the sample size is large enough.

Similar to mSPRT in Section 2.2, if we can show that the ratio $\tilde{\lambda}_n$ is a martingale under null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, the type I error control for SST follows immediately by applying *optional stopping theorem* and the fact that *a mixture of a martingale is also a martingale*. Clearly, as a result of asymptotic distribution and empirical replacement, exact martingale cannot be proved for $\tilde{\lambda}_n$. But with approximate martingale property in Remark 3.1, the decision rule (14) for SST approximately controls type I error at small α where large sample size is necessary to reject H_0.

4 Multiple Testing

The SST framework can also be applied to *multiple testing*, where more than one treatment variation are compared against a baseline variation, or more than one metric are of interest between two variations. The main problem in multiple comparisons is that the probability to find at least one statistically significant effect across a set of tests, even when in fact there is nothing going on, increases with the number of comparisons (Hsu 1996).

In fixed-horizon, *Bonferroni correction* (Miller 1966) and *Benjamini-Hochberg(BH)* (Benjamini and Hochberg 1995) are two well-studied methods designed to address this issue. The Bonferroni correction deals with multiple testing by controlling the *family-wise error rate* (FWER): the probability of making at least one false rejections. Although FWER control provides the safest inference, it is too conservative to offer sufficient detection power. Therefore, the BH procedure is proposed to control the *false discovery rate* (FDR): the expected proportion of the rejections that are false. Both these two procedures take as input the vector of the p-values for each comparison and produce a set of rejections.

In sequential test, the always-valid p-value defined in (2) works as the ordinary p-value in fixed horizon testing. It is trivial to show that Bonferroni or BH procedure applied on a collection of sequential p-values controls FWER or FDR (respectively) in the presence of arbitrary continuous monitoring (Johari et al. 2017). The corresponding algorithms for sequential multiple comparisons under SST framework can be summarized in proposition 1 and 2.

Proposition 1 *(Bonferroni Correction for SST). For arbitrary stopping time T, compute the corresponding sequential p-values $(p_T^i)_{i=1}^m$ by (15) for m comparisons. Then reject hypotheses $(1), ...(j)$, where j is the maximal such that $p_T^{(j)} \leq \alpha/m$, and $p_T^{(1)}, ..., p_T^{(m)}$ are the p-values arranged in an increasing order.*

Proposition 2 *(Benjamini-Hochberg Procedure for SST). For arbitrary stopping time T, compute the corresponding sequential p-values $(p_T^i)_{i=1}^m$ by (15) for m comparisons. Then reject hypotheses $(1), ...(j)$, where j is the maximal such that:*

$$p_T^{(j)} \leq \frac{\alpha j}{m \sum_{r=1}^m 1/r} \tag{23}$$

and $p_T^{(1)}, ..., p_T^{(m)}$ are the p-values arranged in an increasing order.

Note that the term $\sum_{r=1}^m 1/r$ in (23) accounts for the fact that the p-values may be correlated (Benjamini, Yekutieli, and others 2001).

5 Experiment

5.1 Simulation

In this section, we compare our SST with the widely-used mSPRT for both A/B tests (two-variations tests) and multiple tests on simulation data generated from combinations of three generalized linear models (5)-(7)(9) and five types of covariates. The significance level $\alpha = 0.05$, null value of testing parameter $\boldsymbol{\beta}_0 = (0, 0)$ (for 2-dimensional covariates) or $(0, 0, 0)$ (for 3-dimensional covariates) and true nuisance parameter $\boldsymbol{\theta}_0 = (0, 1)$ (2-dimension) or $(0, 1, -1)$ (3-dimension) are fixed for all experiments. Each experiment is repeated 1000 times to estimate type I error and power for SST and mSPRT.

Generalized linear models: We choose three generalized linear models to represent response in different applications. For binary outcomes, such as clicks, conversions, etc., we use logistic regression (Bernoulli distribution). For real-valued response like revenue, ordinary linear regression (normal distribution) is a good choice. If the response are non-negative integers, Poisson distribution which corresponds to log regression is appropriate. However, mSPRT didn't provide the form of test statistics for Poisson distribution, so we only gives our SST result for log regression.

Covariates generation: We consider 5 different distributions for 2 or 3-dimensional ($p = 1$ or 2) covariates. The first dimension is always 1 to indicate the intercept. The other element of 2-dimensional covariates are generated from normal distribution $\mathrm{N}(0, 1)$, uniform distribution $\mathrm{U}[-1, 1]$ and Bernoulli distribution $\mathrm{Ber}(0.5)$, respectively. The last two elements of 3-dimensional covariates are generated either from a multivariate normal with mean $\binom{0}{0}$ and variance

GLM	$\boldsymbol{\beta}_0$	$\boldsymbol{\theta}_0$	Covariates	Type I error (SST)	Type I error (mSPRT)
Logistic Regression	(0,0)	(0,1)	N(0,1)	0.017	0.001
			U[-1,1]	0.019	0.004
			Ber(0.5)	0.016	0.005
	(0,0,0)	(0,1,-1)	MVN	0.023	0.003
			N(0,1)+U[-1,1]	0.026	0.002
Linear Regression	(0,0)	(0,1)	N(0,1)	0.001	0.132
			U[-1,1]	0.003	0.026
			Ber(0.5)	0.005	0.021
	(0,0,0)	(0,1,-1)	MVN	< 0.001	0.136
			N(0,1)+U[-1,1]	< 0.001	0.201
Log Regression	(0,0)	(0,1)	N(0,1)	0.006	NA
			U[-1,1]	0.008	
			Ber(0.5)	0.006	
	(0,0,0)	(0,1,-1)	MVN	0.003	
			N(0,1)+U[-1,1]	0.004	

Table 1: Estimated Type I error for HTE and ATE testing

GLM	$\boldsymbol{\beta}_0$	$\boldsymbol{\theta}_0$	Covariates	Power (SST)	Power (mSPRT)
Logistic Regression	(-0.12,0.12)	(0,1)	N(0,1)	0.730	0.356
			U[-1,1]	0.709	0.514
			Ber(0.5)	0.215	0.079
	(-0.15,0.15)	(0,1)	N(0,1)	0.956	0.655
			U[-1,1]	0.938	0.851
			Ber(0.5)	0.436	0.169
	(-0.12,0.12,-0.12)	(0,1,-1)	MVN	0.544	0.384
			N(0,1)+U[-1,1]	0.559	0.287
	(-0.15,0.15,-0.15)	(0,1,-1)	MVN	0.875	0.636
			N(0,1)+U[-1,1]	0.897	0.564
Linear Regression	(-0.05,0.05)	(0,1)	N(0,1)	0.685	0.607
			U[-1,1]	0.419	0.535
			Ber(0.5)	0.053	0.099
	(-0.08,0.08)	(0,1)	N(0,1)	1	0.943
			U[-1,1]	0.979	0.960
			Ber(0.5)	0.413	0.323
	(-0.05,0.05,-0.05)	(0,1,-1)	MVN	0.400	0.607
			N(0,1)+U[-1,1]	0.602	0.653
	(-0.08,0.08,-0.08)	(0,1,-1)	MVN	0.994	0.955
			N(0,1)+U[-1,1]	0.999	0.943
Log Regression	(-0.05,0.05)	(0,1)	N(0,1)	0.499	NA
			U[-1,1]	0.132	
			Ber(0.5)	0.026	
	(-0.08,0.08)	(0,1)	N(0,1)	0.996	
			U[-1,1]	0.758	
			Ber(0.5)	0.282	
	(-0.05,0.05,-0.05)	(0,1,-1)	MVN	0.242	
			N(0,1)+U[-1,1]	0.726	
	(-0.08,0.08,-0.08)	(0,1,-1)	MVN	0.971	
			N(0,1)+U[-1,1]	1	

Table 2: Estimated power for HTE and ATE testing

$\left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$, or a hybrid distribution with one variable from $\mathrm{N}(0,1)$ and the other one from $\mathrm{U}[-1,1]$ independently.

In A/B testing, data are generated in batch with batch size 200, and then are assigned equally to control group and treatment group. After each batch, we compute the corresponding test statistic and reject the null hypothesis the first time it exceeds some predetermined threshold. We also set an upper bound $N = 10000$, which means that we would accept the null hypothesis if the test statistic does not exceed the threshold before the data are accumulated to $N = 10000$ (for each group). We set the true value of HTE $\boldsymbol{\beta}$ to be 3 vectors with different scales, including the null value $\boldsymbol{\beta}_0$. For $\boldsymbol{\beta} = \boldsymbol{\beta}_0$,

GLM	Covariates	FDR (SST)	FDR (mSPRT)	TPR (SST)	TPR (mSPRT)
Logistic Regression	N(0,1)	0.0119	0.0008	0.8038	0.7191
	U[-1,1]	0.0059	0.0009	0.7957	0.7662
	Ber(0.5)	0.0067	0.0009	0.6501	0.4761
	MVN	0.0114	0.0009	0.7664	0.7171
	N(0,1)+U[-1,1]	0.0148	0.0007	0.7775	0.6944
Linear Regression	N(0,1)	0.0004	0.1983	1	0.3787
	U[-1,1]	0.0009	0.0687	0.9994	0.2868
	Ber(0.5)	0.0011	0.3332	0.8725	0.0504
	MVN	0.0024	0.2013	0.9999	0.3748
	N(0,1)+U[-1,1]	0.0005	0.2708	1	0.4092
Log Regression	N(0,1)	0.0011	NA	1	NA
	U[-1,1]	0.0025		0.9750	
	Ber(0.5)	0.0019		0.8396	
	MVN	0.0010		0.9997	
	N(0,1)+U[-1,1]	0.0011		1	

Table 3: Estimated FDR and TPR of HTE and ATE testing for multiple testing

Control article id	Treatment article id	HTE ($\boldsymbol{\beta}$)	ATE (β)
109510	109520	(-0.401, -0.091, -0.068, 0.661, -0.178)	-0.179

Table 4: Fitted HTE and ATE

we estimate the type I error by computing the rejection ratio among 1000 repeated experiments. For other two vectors, we estimate the power in the same way.

It shows that the sequential score test is able to control type I error (Table 1), and achieve higher detection power (Table 2) than mSPRT if there is heterogeneous treatment effect. We also find that if there exists individual effects on response, that is, $\boldsymbol{\theta} \neq \mathbf{0}$, mSPRT may not be able to control type I error (Table 1). That is because the model assumption for mSPRT given in formula (8) cannot handle individual baseline effects (i.e. $\boldsymbol{\theta}$ in (9)) and possible treatment-covariates interaction effects (i.e. HTE effects described by $\boldsymbol{\beta}$ in formula (9)). Therefore, the mSPRT test cannot adjust baseline covariates and may lead to incorrect type I errors for testing HTE. For example, when $\boldsymbol{\theta} \neq \mathbf{0}$ while $\boldsymbol{\beta} = \mathbf{0}$ in (9), mSPRT may reject the null hypothesis due to the outcome difference caused by baseline effects, which may lead to inflated type I errors. On the other hand, when $\boldsymbol{\beta} \neq \mathbf{0}$, the mSPRT may fail to detect the HTE (lose power) or need to wait a long time to reject the null hypothesis since treatment effects may be masked by individual heterogeneity.

Our proposed test also works with high-dimensional covariates. We conduct additional simulations with 21 covariates (p=20) under logistic regression. Except the first dimension (being 1), the last 20 covariates are independently generated from different distributions. Among these 20 covariates, 7 are generated from normal distribution with variance 1 and different means between -0.3 to 0.3, 8 covariates are from uniform distribution with mean 0 and upper limit between 0.3 to 1, and the last 5 covariates are generated from binomial distribution with probability between 0.1 to 0.5. The individual baseline effect $\boldsymbol{\theta}$ has two non-zero components. The simulation result shows that our SST still can control type I error under the null (type I error is 0.025, which is less than the significance level α), and has reasonable power (i.e. when HTE effects $\boldsymbol{\beta}$ has three non-zero components with the value of 0.2, the power is 0.731; and when $\boldsymbol{\beta}$ has three non-zero components with the value of 0.3, the power is 1).

In multiple testing, the configurations of the hypotheses involve $m = 64$ hypotheses, $\frac{3}{4}m$ true null hypotheses ($\boldsymbol{\beta}_0 = (0,0)$ or $(0,0,0)$) and the remaining $\frac{1}{4}m$ true alternatives being equally placed at $\boldsymbol{\beta}_0 = (-B, B)$ or $(-B, B, -B)$, where $B = 0.1, 0.2, 0.3, 0.4$, respectively. For each comparison, we wait until the data are accumulated to $N = 10000$ (for each group) and compute the sequential p-value p_N according to (15). After applying *Benjamini-Hochberg(BH)* procedure, we get the rejections from which we can estimate FDR and *true positive rate* (TPR), also known as *recall*. The TPR, defined as the proportion of correctly rejections in truly alternatives, is a metric for detection power in multiple testing. Same as A/B testing, the results (Table 3) show that SST applied on multiple testing achieves higher TPR than mSPRT while maintaining FDR in control.

5.2 Real Data

We also compare SST with mSPRT on Yahoo dataset which contains user click events on articles over 10 days. Each event has a timestamp, a unique article id, a binary click indicator, and five user features which are between 0 and 1 and sum to 1 for each user (we only use the last four features). We treat each article as different treatment variations, click actions as the binary responses. Our goal is to test if there is any article effects on user click behaviors with (SST) or without (mSPRT) accounting for the user features.

We first conduct A/A test to show the validity of test on click events with the most popular article (id=109510) on the date May 1st, 2009, by randomly assigning fake treat-

ment indicators to them. Then we conduct A/B test on events with two most popular articles (id=109510 and 109520) on the date May 1st, 2009. With every 200 events (from both articles) coming in a time sequence, we compute the corresponding test statistics. As soon as the statistics exceed the predetermined critical value $(1/\alpha)$, we stop and reject the null hypothesis. If all the data are used up, we accept the null hypothesis. The experiment shows that both SST and mSPRT accept the null hypothesis for A/A test, indicating type I errors are well controlled for both tests under the considered hypotheses. For the A/B test, SST needs $n = 19600$ events to get rejection conclusion while mSPRT needs $n = 67600$. It means that we are able to discover the difference early by accounting for the covariates. We also provide estimated HTE ($\boldsymbol{\beta}$ in (9)) and ATE (β in (8)) by fitting logistic regression.

For multiple test, we choose 10 articles and do pairwise comparisons. Hence, there are $m = 45$ comparisons in total. We compute p_T for each pair with $T = 20000$ from each article and then apply BH procedure. Among 45 pair comparisons, we reject 43 with SST and 23 with mSPRT.

6 Conclusions

We propose a new framework of online test based on the probability ratio of score function. It is able to test a multi-dimensional heterogeneous treatment effect while accounting for the unknown individual effect. The asymptotic normality of the score function guarantees an explicit form, greatly improving the computation efficiency. We provide an online p-value for SST and extend the procedure to online multiple testing. We validate our testing procedure by both theoretical proof and empirical results. We also compare it with a state-of-art online test named mSPRT on simulation and real data. The results show that our proposed test controls type I error at any time, has higher detection power and allows quick inference on online A/B testing.

There is still some interesting work we may do in the future. The decision rule of our test implies that we can only get rejection conclusions unless we wait essentially indefinitely, which is impossible in practice. This necessitates truncating SST at a maximum size and admitting an inclusive result if we ever reach it, which may diminish the power more or less. How to choose the truncating size to trade off between waiting time and power still remains a problem.

References

Balsubramani, A., and Ramdas, A. 2015. Sequential nonparametric testing with the law of the iterated logarithm. *arXiv preprint arXiv:1506.03486*.

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.; Yekutieli, D.; et al. 2001. The control of the false discovery rate in multiple testing under dependency. *The annals of statistics* 29(4):1165–1188.

Goodson, M. 2014. Most winning a/b test results are illusory. *Whitepaper, Qubit, Jan*.

Grimmer, J.; Messing, S.; and Westwood, S. J. 2017. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis* 25(4):413–434.

Grimmett, G.; Grimmett, G. R.; and Stirzaker, D. 2001. *Probability and random processes*. Oxford university press.

Hsu, J. 1996. *Multiple comparisons: theory and methods*. Chapman and Hall/CRC.

Johari, R.; Koomen, P.; Pekelis, L.; and Walsh, D. 2017. Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1517–1525. ACM.

Johari, R.; Pekelis, L.; and Walsh, D. J. 2015. Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922*.

Ju, N.; Hu, D.; Henderson, A.; and Hong, L. 2019. A sequential test for selecting the better variant: Online a/b testing, adaptive allocation, and continuous monitoring. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 492–500. ACM.

Kohavi, R.; Longbotham, R.; Sommerfield, D.; and Henne, R. M. 2009. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery* 18(1):140–181.

Kulldorff, M.; Davis, R. L.; Kolczak, M.; Lewis, E.; Lieu, T.; and Platt, R. 2011. A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Sequential analysis* 30(1):58–78.

Lai, T. L. 2001. Sequential analysis: some classical problems and new challenges. *Statistica Sinica* 303–351.

Lehmann, E. L., and Romano, J. P. 2006. *Testing statistical hypotheses*. Springer Science & Business Media.

Malek, A.; Katariya, S.; Chow, Y.; and Ghavamzadeh, M. 2017. Sequential multiple hypothesis testing with type i error control. In *Artificial Intelligence and Statistics*, 1468–1476.

Miller, R.G., J. 1966. *Simultaneous Statistical Inference*. New York: McGraw-Hill Book Co.

Pekelis, L.; Walsh, D.; and Johari, R. 2015. The new stats engine. *Internet. Retrieved December* 6:2015.

Pollak, M. 1978. Optimality and almost optimality of mixture stopping rules. *The Annals of Statistics* 910–916.

Robbins, H., and Siegmund, D. 1974. The expected sample size of some tests of power one. *The Annals of Statistics* 415–436.

Robbins, H. 1970. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics* 41(5):1397–1409.

Simmons, J. P.; Nelson, L. D.; and Simonsohn, U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 22(11):1359–1366.

Wald, A. 1945. Sequential tests of statistical hypotheses. *The annals of mathematical statistics* 16(2):117–186.