Objective Bayesian Two Sample Hypothesis Testing for Online Controlled Experiments

Alex Deng
Microsoft
One Microsoft Way
Redmond, WA 98052
alexdeng@microsoft.com

ABSTRACT

As A/B testing gains wider adoption in the industry, more people begin to realize the limitations of the traditional frequentist null hypothesis statistical testing (NHST). The large number of search results for the query "Bayesian A/B testing" shows just how much the interest in the Bayesian perspective is growing. In recent years there are also voices arguing that Bayesian A/B testing should replace frequentist NHST and is strictly superior in all aspects. Our goal here is to clarify the myth by looking at both advantages and issues of Bayesian methods. In particular, we propose an objective Bayesian A/B testing framework for which we hope to bring the best from Bayesian and frequentist methods together. Unlike traditional methods, this method requires the existence of historical A/B test data to objectively learn a prior. We have successfully applied this method to Bing, using thousands of experiments to establish the priors.

Category and Subject Descriptors: G.3 [Probability and Statistics]: Statistical Computing

Keywords: A/B testing, controlled experiments, Bayesian statistics, prior, objective Bayes, empirical Bayes, multiple testing, optional stopping

1. INTRODUCTION

The last decade witnessed a strong revival of Bayesian methods, dating back over 250 years. It also witnessed the trend of two sample hypothesis testing, a century old frequentist statistical method [32] originally designed for small data sets, now being applied to terabytes of data collected online with extremely low cost [19; 33]. The application of two sample hypothesis testing, often under the alternative name of A/B testing in the industry, has been established as a cornerstone for a data-driven decision making culture [21; 20; 18; 6; 8]. It's not surprising that there is a growing interest in applying Bayesian method in A/B testing. In fact, a web search of the query "Bayesian A/B testing" returns millions of results. [1]

In those articles or blogs freely accessible from the Internet(also journal publications such as Kruschke [22]), Bayesian framework is often pictured as strictly superior to the traditional frequentist null hypothesis statistical testing (NHST). Some people even claim that everybody conducting A/B testing should be taking the Bayesian approach and the reason that Bayesian method has long been shadowed by frequentist method is solely because of the lack of computational power which is now largely irrelevant with efficient Markov Chain Monte Carlo (MCMC). In the peer reviewed statistics literature, the debate between the two schools also never ended and many great statisticians have contributed to the discussion and searched for a common ground where frequentists and Bayesians can make peace and even join force with synergy [9; 12; 2; 3]. Although there is a deep philosophical component of the fight between Bayesian and frequentist, many recent researches have been taking a pragmatic perspective, showing nice frequentist properties of Bayesian methods and many frequentist methods have a Bayesian perspective. In particular, regularization methods that are widely used in machine learning [26], mixed effect model or multi-level modeling [15] and false discovery rate control [1; 10; 24] all can be seen as examples of Bayesian-frequentist synergy. Many recent works focused on bringing scientific objectiveness to Bayesian method, so that it can be accepted as a scientific standard [2; 14; 3; 13].

This paper follows the same line of thinking pursuing objectiveness of Bayesian A/B testing. We introduce readers to the Bayesian way of hypothesis testing and explain why it is indeed a conceptually unified and elegant framework that avoids many issues its frequentist counterpart faces such as multiple testing and optional stopping when the true prior is known or can be learned with enough accuracy. The main part of this paper considers an objective Bayesian two sample hypothesis testing procedure learning its prior through prior experiments. We also illustrate why Bayesian approaches made popular online in aforementioned blogs are often limited or impractical.

The main contribution of this paper is the objective Bayesian hypothesis testing framework in Section 3. This framework provides a solution for multiple testing, optional stop-

[1]For instance, blog posts from www.bayesianwitch.com, Lyst's developer blog developers.lyst.com, Evan Miller's closed-form formula for posterior probability of a lifted conversion rate under the Beta-Bernoulli model. Interestingly but not surprisingly, there are more search results for query "Bayesian A/B testing" than "Bayesian hypothesis testing".

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742563.

ping, power analysis and metric sensitivity analysis. Simulation and empirical results are also presented in Section 4 and 5. We hope this paper can point researchers and practitioners in this big data era to a new direction of objective Bayesian analysis where it is possible to use historically collected data to avoid the subjectiveness of Bayesian modeling.

We assume readers are familiar with the concepts of null hypothesis statistical testing in controlled experiments and familiar with the Bayes Rule. Readers new to these concepts should refer to references such as Kohavi et al. [21].

2. BAYESIAN ADVANTAGES AND ISSUES

We illustrate Bayesian hypothesis testing using a simple example from Efron [12] and leave the formulation of Bayesian two sample hypothesis testing for Section 3. A physicist found out she was going to have twin boys and wanted to know the chance of the twins being *Identical* rather than *Fraternal*. According to the doctor, *past experience* favors *Fraternal* with the prior odds between these two hypotheses being

$$\frac{P(\text{identical})}{P(\text{fraternal})} = \frac{1/3}{2/3} = \frac{1}{2} \quad \text{(Prior Odds)}.$$

On the other hand, given the observed data that the twins being twin boys,

$$\frac{P(\text{twin boys}|\text{identical})}{P(\text{twin boys}|\text{fraternal})} = \frac{1/2}{1/4} = 2 \quad \text{(Bayes Factor)},$$

where the first equality is based on the belief that boys and girls have equal probability and the fact that identical twins must have the same sex while fraternal sexes are result of two independent fair coin flipping. This ratio represents the odds of the two hypotheses based on evidence from the data. The numerator and denominator is the likelihood of observing the data under the identical twin hypothesis and the fraternal twin hypothesis, respectively. It is also called Bayes factor [17].

The last step applies the Bayes Rule by combining the two pieces of information to get the posterior odds.

$$\text{Posterior Odds} = \text{Prior Odds} \times \text{Bayes Factor}. \quad (1)$$

In this case the posterior odds is $1/2 \times 2 = 1$. Therefore a Bayesian would decide the chance of *Identical* or *Fraternal* to be 50%.

The twins problem has a simple structure. But it captures the essence of Bayesian Model Comparison or Bayesian Hypothesis Testing. Under the Bayesian framework, when trying to guess the true state from two possible states, one always need two pieces of information: the prior odds and the likelihood ratio. The former represents prior belief while the latter represents the evidence manifested through data.

The twins problem also sets an example that a frequentist will likely also agree with the deduction. To see why, the frequentist and the Bayesian only needs to agree upon the prior odds. The frequentist might question the source of this *prior experience* and learns that the doctor, having observed birth rate of twins, was giving this prior odds based on historical birth rate. The frequentist might also want to get the data from which the doctor got his estimate but he would certainly not fight against the Bayesian because he knows he can study the accuracy of this procedure given the birth rate data under the frequentist framework and the result would likely to be the same if the birth data is large enough. Following frequentist philosophy, we call this type of Bayesian analysis *objective* because conclusions (e.g Type-I error, confidence interval coverage) can be verified by a thought experiment where we can repeat the procedure independently for many times [2].

There will be peace between Bayesian and frequentist if all problems are this simple. The prior odds in the twins problem is an example of *genuine prior* [12] and the Bayesian/frequentist controversy centers on the use of Bayes rule in the absence of such prior. Cases like this are not just toy examples. We will show that with hundreds or thousands of *historical experiments* available, just like having birth rate data as side information, the two sample hypothesis testing problem can be very similar to the twins problem and it is possible for us to have an objective Bayesian testing procedure.

Leaving the objectiveness of Bayesian procedure for later discussion, assuming the genuine prior is known, we now give a few reasons why a Bayesian procedure conceptually could be a better alternative than frequentist null hypothesis statistical testing (NHST). Many issues of frequentist statistics have been discussed elsewhere, see Murphy [25, Section 6.6].

Straightforward Interpretation and Unified Framework

Frequentist concepts such as p-value and confidence interval are often criticized by their unintuitive definitions. In our experience, many people, especially engineers and business executives often misinterpret p-value as $P(H_0|\text{Data})$ — Bayesian posterior of the null hypothesis. We believe this reflects the fact that people often find Bayesian posterior a more natural way of summarizing strength of the statistical evidence.

Bayesian reasoning is also a unified framework based on the Bayes rule. All reasoning center around updating prior into posterior given data. This is in stark contrast to frequentist methods where different problems need to be tackled using different methods.

Accumulative Evidence and Optional Stopping

Bayesian methods naturally handle evidence update. It is straightforward to combine evidence from repeated tests using belief update, in contrasts to frequentist approaches using meta-analysis or multiple testing adjustment [7].

As noted in Berger and Bayarri [3] and Berger and Wolpert [4] as the *stopping rule principal*, Bayesian methods with a genuine prior automatically support optional stopping, i.e. the experimenter can choose to stop collecting data once the posterior evidence is strong enough. Also see Rouder [27] for related debate in the literature and some simulation justifications. On the contrary, NHST is incompatible with optional sampling unless special sequential or group sequential test procedure is followed.

Multiple Testing

Somewhat related to optional stopping, multiple testing is another issue in NHST. The root cause of the problem is connected to the way null hypothesis testing is constructed to control Type-I error. Multiplicity will inflate the Type-I error. Traditional frequentist methods struggle to control FWER (family-wise error rate), e.g. Bonferroni correction, but typically found the criteria too restrictive. Modern multiple testing procedure are all based on the notion of false discovery rate (FDR), which resembles the Bayesian posterior $P(H_0|\text{Data})$. See Benjamini and Hochberg [1]; Efron [10] and Muller et al. [24].

There is no coincidence that FDR resembles Bayesian posterior and many frequentist works in this area borrow ideas from Bayesian methods. Bayesian reasoning with a genuine prior automatically adjusts for multiplicity in many cases [11; 31]. Take the twin's problem again for illustration, if we have another 100 or 1,000 twins for whom we need to test the hypothesis of identical vs. fraternal, assuming no connection between those twins whatsoever, after getting the posterior probabilities for each and every one of them, our job is done. There is no need for multiple testing adjustment because for any given twin, all observations for other twins should not interfere with reasoning for this given twin.

Accepting the Null

Another issue of frequentist NHST is that it is not consistent in the sense that chance of accepting the null hypothesis does not converge to 100% with infinite data when the null is indeed true. To see this, p-value under mild regularization condition will follow uniform(0,1) distribution under the null hypothesis no matter how many data are observed. Hence the Type-I error stays at 5% when 5% is the chosen significant size. NHST is just not designed to accept the null hypothesis.

2.1 Importance of a Genuine Prior

Bayesian methods have many advantages beyond what we listed above, such as dealing with nuisance parameter, etc. However, not everyone is a Bayesian. Instead, frequentist NHST is still the dominating standard for scientific discovery. The main reason is the choice of prior. Typically, one either choose a prior from a distribution family that is easier for computation (conjugate prior), or use so-called "uninformative" priors. First, there is no evidence that the prior should come from the conjugate family. Secondly, there is no truly uninformative prior and every prior carries information. For example, assigning uniform prior to a parameter gives information of its range, and an improper uniform prior would suggest this parameter could be very large and would cause a well-known problem called "Lindley's paradox" [25]. Jeffery's prior, another common choice of uninformative prior, only makes sure the inference won't be affected by transformation of the prior. In literature, almost all Bayesian approaches nowadays chose certain uninformative prior since computational cost is lower thanks to procedures like MCMC. However, this is far from saying that the choice of prior is genuine like in the Twin's problem.

Why is a genuine prior so important? First it is obvious that scientific objectiveness is crucial for any method to be set as a standard. For A/B testing one might be willing to take a more pragmatic view since after all the ultimate goal is not to submit the result for the scientific community to review, and the benefits from using Bayesian approach as listed above seem to strongly prefer using Bayesian method without the need of a genuine prior. However, another important point is that many benefits we listed above relies on the prior to be a genuine prior, not just any prior. In fact, optional stopping and multiple testing pose no issue for Bayesian methods only when the right prior is being used! One way to see why is to notice that in Bayesian posterior estimation, assuming uniform prior on a parameter, maximizing the posterior is mathematically equivalent to maximizing the likelihood, and the posterior mode is the MLE. On the other hand, when a non-uniform prior is applied, the posterior mode can be seen as a smoothed or regularized version of the MLE. From this perspective, knowing the genuine prior as in the twin's problem is like applying the right amount of smoothing such that the estimation is neither too aggressive nor too conservative. Without knowing the right amount of smoothing, we can either be too aggressive like using uniform prior and therefore falling into the trap of multiple testing and optional stopping, or be too conservative.

3. OBJECTIVE BAYESIAN A/B TESTING

In this section we propose an objective Bayesian A/B testing procedure where we try to model and fit the genuine prior objectively from historical data. Suppose the data we observed for treatment and control groups are i.i.d. observations from two distributions with unknown mean τ_T and τ_C respectively. Denote our observations by $Y_i, i = 1, \ldots, N_T$ and $X_i, i = 1, \ldots, N_C$. We want to test the null hypothesis $H_0 : \tau_T - \tau_C = 0$ against the alternative $H_1 : \tau_T \neq \tau_C$.

Without assuming distributions of X and Y, in A/B testing we normally resort to the central limit theorem and hence use Wald test which can be seen as large sample version of the well-known t-test. The test statistic is

$$\mathrm{Z} := \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_T^2/N_T + \sigma_C^2/N_C}} = \frac{\Delta}{\sqrt{\sigma_T^2/N_T + \sigma_C^2/N_C}},$$

where σ_C and σ_T are variances of X and Y. The variances are also unknown but in large sample scenario we can assume they are known and use their estimates. This is because the test is already using asymptotic normal approximation. Note that metrics are often in different scales. We first define $N_E = 1/(1/N_T + 1/N_C)$ to be the **effective sample size**. And then let σ^2 be the **pooled variance** such that $\sigma^2/N_E = \sigma_T^2/N_T + \sigma_C^2/N_C$. With $\delta = \Delta/\sigma$, Z-statistics can be rewritten as

$$\mathrm{Z} = \frac{\delta}{\sqrt{\sigma^2/N_E}}. \tag{2}$$

δ is Δ scaled by pooled standard deviation and is called the **effect size**. Finally, define

$$\mu := E(\delta) = E(\Delta)/\sigma = (\tau_T - \tau_C)/\sigma \tag{3}$$

is the average treatment effect scaled by σ. When σ is known, inference on $\tau_T - \tau_C$ and μ are equivalent. In Bayesian analysis it is common to define prior for μ as it is scaleless.

3.1 One Group and Two Group Model

There are in general two approaches for two sample hypothesis testing problem in Bayesian literature and here we call these two one group model and two group model.

In one group model, we observe $\delta \sim N(\mu, 1/N_E)$ where μ has a prior distribution with density $\pi(\mu)$. Under this model, we then focus on inferring the posterior distribution of δ given the observation δ. See, for example, [22] and [11]. Notice there is no special mentioning of H_0 and H_1 in this model and in particular if the prior π has continuous density, then $P(\mu|\delta)$ is also continuous. Hence it is inevitable that we have $P(H_0|\delta) = P(\mu = 0|\delta) = 0$. We can avoid this issue by assuming there is a region of practical equivalence (ROPE) around 0 that we can define as H_0 so $P(H_0|\delta)$ is nonzero. Another approach is to put a nonzero probability

in 0 on the prior π itself, making it not continuous at 0, as in the two group model.[2]

A two group model assumes two different priors for H_0 and H_1. Under the null H_0, $\mu = 0$. Under the alternative H_1, we assume a prior π for μ. For both cases we observe $\delta \sim N(\mu, 1/N_E)$. In addition, we assume a prior probability p for H_1 being true. The goal of the analysis is to infer the posterior $P(H_0|\delta)$ as well as the posterior distribution of μ. This model is very similar to the twin's problem except that we have one extra prior distribution π for μ under H_1. Posterior $P(H_1|\delta)$ can also be inferred by prior odds $\times$ likelihood ratio = posterior odds.

There are no essential differences between two models and the two group model can be seen as one group model with a pre-specified nonzero probability at 0. The key issue here is how can we achieve objectiveness, i.e. how can we choose the prior π and p without assuming too much. In the following we use the two group model. The method we described here can also be used in the one group model with or without ROPE. Results under the one group model are left as future work.

3.2 Learning the Objective Prior

Literature using the two group models often avoids discussing the choice of prior probability p of H_1 being true by assuming prior odds of H_1 against H_0 to be 1 or simply leave the decision to the users. Methods proposed differ in the way prior distribution $\pi(\mu)$ is chosen. One of them is the unit-information prior where $\mu \sim N(0, 1)$. Another choice closely related to unit-information prior is using Bayesian information criterion [23; 17], which is based on a rough Laplace approximation to the log Bayes factor. Rouder et al. [28] proposed JZS prior and compared it to other priors aforementioned.

Here we take advantages of historical experiment results and use them to learn the prior. Suppose for a given metric, we have N previously conducted tests with observed effect size and effective sample size $(\delta_i, N_{Ei}), i = 1, \ldots, N$. We have no idea which of those are from H_0 or H_1. Next we put π into a parametric family such as exponential family[11]. The two group model, with a parametric model of π and prior probability p, formed a generative model for δ_i and we can estimate model parameters of π and p using maximum likelihood. In fact, this approach is called *Empirical Bayes* or ML-II[13; 25].

Although an exponential family model for $\pi(\mu)$ up to some degree of freedom is more general, here we use a simple $N(0, V^2)$ model. The reason being: 1) the model is simple and relatively easy to fit. It is a special case of exponential family up to the 2nd degree and too large degree of freedom requires larger N to make the estimation robust; 2) the result is also easier to interpret and illustrate with limited space in this paper. The idea here extends to the general case.

Fitting the model to find MLE isn't straightforward, due to the fact that we don't know each δ_i belongs to H_0 or H_1. Fortunately, a solution for this type of hidden latent variable problem, called Expectation-Maximization, is well-known[5]. EM algorithm in our case reduces to a fairly intuitive form as the following.

Step I. If p and V are known, the posterior odds for each δ_i belonging to H_1 against H_0 have the simple form

$$\frac{\phi(\delta_i; 0, 1/N_{Ei} + V^2)}{\phi(\delta_i; 0, 1/N_{Ei})} \times \frac{p}{1-p} \tag{4}$$

where $\phi(x; \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2. Convert posterior odds to $P_i := P(H_1|\delta_i; p, V)$.

Step II. Set p to be $\overline{P(H_1|\delta; p, V)}$ by taking average of all P_i calculated in Step I.

Step III. To update V, note that under H_1, $Var(\delta_i) = E(\delta_i^2) = V^2 + E(1/N_{Ei})$. Although we don't know absolutely whether a δ_i belongs to H_1, we can use posterior P_i in Step II as weights:

$$V^2 = \text{WAvg}(\delta_i^2; P_i) - \text{WAvg}(1/N_{Ei}; P_i) \tag{5}$$

where $\text{WAvg}(x_i; w_i) = \sum w_i x_i / \sum w_i$. To avoid numerical issue that V^2 in (5) can take negative value, we bound V^2 away from 0 by a lower bound.

The EM algorithm starts with an initial value of p and V, iterates through the 3 steps above until they converge. Step I is the E-step. Step II and Step III are the M-step updating p and V.(Technically Step III is generalized M-step. We update V using method of moment estimator knowing with high probability it increase the expected log-likelihood.) The algorithm is summarized in Algorithm 1. We can also learn parameters of π and p using a full (Hierarchical) Bayesian approach by putting another layer of prior for p and V as in Scott and Berger [29]. The comparison of Empirical Bayesian and Full Bayesian is beyond the scope of this paper.

Algorithm 1 Fitting p and V with EM algorithm

```
Initialize p and V
repeat
    Step I: Update P_i using (4)
    Step II: p ← mean(P_i)
    Step III: Update V using (5)
until Converge
```

The lower bound in Step III is not purely a numerical trick. It is needed for model identification. When $V = 0$, $\mu \equiv 0$ under both H_1 and H_0. We cannot distinguish H_1 and H_0, leading to an unidentifiable model. We recommend setting the lower bound $V^2 = k^2 \times \text{Avg}(1/N_E)$ and set k to 2, for reason which will be make clear below.

3.3 Practical Applications

We have proposed algorithm 1 to learn a prior directly from historical data. In this section we talk about practical applications. As laid out in Section 2 we are now in the position of reaping many Bayesian benefits. On top of the list is automatic multiple testing adjustment and the ability to do optional stopping. Our framework also enables us to combine results from repeated experiments by straightforward Bayesian belief update, avoiding techniques as in Deng et al. [7]. We will also accept the null given enough data if null is indeed true, achieving consistency. Here we mention 3 other applications useful in A/B testing.

P-Assessment. The concept of P-Assessment is an application of posterior probability. Instead of reporting p-value, we can now report three probabilities, P(Negative), P(Flat)

[2]It is true that the distribution of μ contains more information than a value like $P(H_0|\delta)$ so we don't necessarily require a nonzero $P(H_0|\delta)$. However having a nonzero $P(H_0|\delta)$ makes a more direct comparison to to frequentist A/B testing.

and P(Positive), defined as the posterior probability of μ being negative, 0 and positive given the data. P-Assessment is more intuitive than p-value, and also separates positive and negative movements. If we intend to ship a feature if there is no degradation, e.g. a code clean-up, we can look at P(Negative) to see whether it is small enough for us to be confident enough in the no degradation claim. If we intend to ship a feature if there is an improvement, we then look at P(Positive). For a metric, given its learned prior parameters p and V as well as observed δ and effective sample size N_E. We first use (4) to get posterior odds and then convert to $P(H_0|\text{Data})$. Under H_1, the posterior of μ is $N(A\delta, A/N_E)$ where $A = V^2/(V^2 + 1/N_E)$ is the shrinkage factor. Then $\text{P(Flat)} = P(H_0|\text{Data})$ and

$$\text{P(Negative)} = (1 - P(Flat)) \times \Phi(0; A\delta, A/N_E),$$
$$\text{P(Positive)} = 1 - P(Flat) - \text{P(Negative)},$$

where $\Phi(x; \theta, \sigma^2)$ is the normal cumulative density function.

Power Analysis. Power in NHST depends on the unknown true alternative. In the Bayesian framework since we model the alternative prior as $N(0, V^2)$ we can calculate the marginal power. Let $V^2 = k^2 \times 1/N_E$. It can be shown that the power is $2(1-\Phi(1.96; 0, 1+k^2))$. When $k = 2$, the marginal power is 38%. This justifies the choice of k for the lower bound of V, i.e., we assume the alternative at least provides a marginal power of about 40%.

Metric Sensitivity Evaluation. In A/B testing we often tweak metric definitions. A common task is to improve metric sensitivity. It is made clear in the two group model that metric sensitivity has two separate components. One is whether the features we came up with truly moved the metric, represented by prior parameter p. The other is the power, i.e., if the metric moved, can we detect it. The latter can be compared by k.

3.4 Problems of Beta-Bernoulli Model

Almost all search results using the query "Bayesian A/B Testing" focus on a particular metric called conversion rate, or simply a rate metric, defined as the count of success divided by number of trials. It's true that conversion rate is the most widely used metrics in A/B testing. However many metrics we care do not fall into this category, such as revenue per user and page loading time. This is why the Gaussian model above is more flexible as it deals with all kinds of metrics as long as central limit theorem applies. Nevertheless, limited to conversion rate is not the most critical issue.

For conversion rate, people often think about Bernoulli or Binomial distribution. Under this model, each page-view is an independent trial and each conversion counts as a success, we want to infer the success rate p. In particular, for A/B testing, we want to compare p_1 and p_2 from treatment and control. For computational simplicity, a common beta prior distribution for both treatment and control is used and closed-form formula for posterior $P(p_1 > p_2|Data)$ is known.

There are at least two issues regarding the method popular online. First is the sampling distribution. Assuming page-view to be independent usually requires a page level randomization. This is different from most applications of A/B testing where user or surrogate like user cookie is used for randomization. The reason is obvious in that we don't want user to experience switching experience and also for the purpose of being able to track user metrics such as active days per user. When user is used as the randomization unit, the Bernoulli distribution assumption, and hence the whole Beta-Bernoulli model should not be used as it underestimates the variance and Type-I error by assuming too much independence.

Secondly, even if page-view is used as the randomization unit, the prior used is not genuine. In particular, since the Beta prior can be interpreted as trials and success data collected prior to the experiment, the posterior mean of p is a smoothed version of MLE and the posterior of $p_1 - p_2$ is shrunk towards 0. It is often presented as a hint that multiple testing and optional stopping can be used with this kind of Bayesian method. As we explained in Section 2 that Bayesian method provides a hope to avoid the need of multiple testing adjustment and allows optional stopping. However, all these promises depend on knowing the genuine prior (Section 2.1). Also see Scott and Berger [29] where the authors emphasized the importance of using a prior learned from data to avoid multiple testing adjustment. To the time of writing this paper, we didn't find any online blogs or articles discussed the procedure of objectively learning prior.

4. SIMULATION RESULTS

To illustrate consistency of Algorithm 1, we use simulation study. First, we show that if the prior $P(H_0)$ is 100%, our method will uncover this. We varied the number of historical data points N from 100 to 2,000. In each case we fix the effective sample size N_E to be $1E6$ so $\delta \sim N(0, 1E-6)$. Table 1 reported $\widehat{P(H_0)} = 1 - \widehat{p}$, and its standard deviation(estimated with 10,000 bootstrap) in parenthesis. We see that except when $N = 100$, we correctly uncovered $P(H_0)$ is 1. Even for the case $N = 100$ the estimation is not very off.

N	100	200	1,000	2,000
$\widehat{P(H_0)}$	0.987(0.040)	1.000(0.0007)	1.000(0.004)	1.000(0.0005)

Table 1: Simulation Results when $P(H_0) = 1$.

Next we simulated mixtures of H_0 and H_1. We varied $P(H_0)$ from 95% to 50%. In each case, we also varied V by using different k ($V^2 = k^2 \times 1/N_E$). Intuitively we expect smaller relative standard deviation for $\widehat{V}$ as $P(H_0)$ decrease since more cases are under H_1 for us to better estimate V. Also, the larger the V, the easier to separate H_0 with H_1 so we expect standard deviation of $\widehat{P(H_0)}$ to decrease too. Both intuition are verified in Table 2 and Table 3 showing results when $N = 2,000$ and $N = 200$ respectively. When $N = 2,000$, we found both p and V can be estimated with reasonable accuracy. Accuracy of V is worse than p, especially when $P(H_0)$ is high. When $N = 200$, without surprise we found accuracy degraded comparing to $N = 2,000$. However the accuracy of p and V are still close enough to the ground truth for this method to be applied in practice, depending on how sensitive the posterior is to the variation of p and V.

5. EMPIRICAL RESULTS

We also applied Algorithm 1 using Bing experiments data. After data quality check, we found for many metrics except a few recently added ones, we typically had more than 2,000 historical data. After fitting the model, we found the prior

N=2000		k=4(V=4E-3)	k=8(V=8E-3)	k=10(V=1E-2)
$P(H_0) = 95\%$	$\widehat{P(H_0)}$	0.962(0.008)	0.953(0.006)	0.953(0.006)
	$\widehat{V}$	4.23E-3(0.55E-3)	7.76E-3(0.82E-3)	9.66E-3(0.98E-3)
$P(H_0) = 90\%$	$\widehat{P(H_0)}$	0.909(0.012)	0.903(0.009)	0.903(0.008)
	$\widehat{V}$	3.81E-3(0.31E-3)	7.42E-3(0.48E-3)	9.28E-3(0.58E-3)
$P(H_0) = 80\%$	$\widehat{P(H_0)}$	0.798(0.015)	0.802(0.011)	0.802(0.011)
	$\widehat{V}$	3.89E-3(0.18E-3)	7.83E-3(0.31E-3)	9.80E-3(0.37E-3)
$P(H_0) = 50\%$	$\widehat{P(H_0)}$	0.487(0.020)	0.491(0.015)	0.492(0.014)
	$\widehat{V}$	3.88E-3(0.11E-3)	7.77E-3(0.19E-3)	9.73E-3(0.23E-3)

Table 2: Mixture of H_1 and H_0. N = 2,000.

N=200		k=4(V=4E-3)	k=8(V=8E-3)	k=10(V=1E-2)
$P(H_0) = 95\%$	$\widehat{P(H_0)}$	0.965(0.019)	0.963(0.016)	0.962(0.016)
	$\widehat{V}$	4.44E-3(1.04E-3)	8.67E-3(2.04E-3)	1.07E-2(0.25E-2)
$P(H_0) = 90\%$	$\widehat{P(H_0)}$	0.925(0.034)	0.907(0.026)	0.908(0.025)
	$\widehat{V}$	3.62E-3(0.75E-3)	6.80E-3(1.12E-3)	8.55E-3(1.45E-3)
$P(H_0) = 80\%$	$\widehat{P(H_0)}$	0.869(0.042)	0.843(0.033)	0.835(0.033)
	$\widehat{V}$	3.94E-3(0.69E-3)	7.40E-3(1.10)E-3	9.05E-3(1.33E-3)
$P(H_0) = 50\%$	$\widehat{P(H_0)}$	0.594(0.067)	0.518(0.051)	0.506(0.047)
	$\widehat{V}$	3.44E-3(0.37E-3)	6.42E-3(0.59E-3)	7.94E-3(0.71E-3)

Table 3: Mixture of H_1 and H_0. N = 200.

$P(H_1) = p$ ranges from as much as 70% to less than 1%. The ordering of those p for different metrics aligns well with our perception of how frequently we believed a metric truly moved. For example metrics like page loading time moved much more often than user engagement metrics such as visits per user. For most metrics p is below 20%. This is because the scale of Bing experimentation allows us to test more aggressively with ideas of low success rate. We also used P(Flat) in the P-Assessment and only looked at metrics with P(Flat)<20% and found it very effective in controlling FDR. Compared to other FDR method [1; 24], our method is the first that takes advantages of metric specific prior information.

6. CONCLUSION AND FUTURE WORKS

In this paper we proposed an objective Bayesian A/B testing framework. This framework is applicable when hundreds or thousands of historical experiment results are available, which we hope will be soon common in this big data era. An natural and important question is how to pick such a set of historical experiments. In principle, when analyzing a new experiment, we want to use only *similar* historical experiments for prior learning. Similarity can be judged by product area, feature team and other side information. However, if we put too many selecting criteria, we will eventually face the problem of not having enough number of historical experiments for an accurate prior estimation, similar to the cold-start problem. One solution is to use the prior learned from all the experiments as a baseline global prior so the prior for a subtype of experiment is a weighted combination of this global prior and the prior learned from the (possibly small) restricted set of historical data. This can be done via hierarchical Bayes, i.e. putting a prior on prior. Other future works include using more general exponential family for $\pi(\mu)$, and also using one group model as in [11].

References

[1] Benjamini, Y. and Hochberg, Y. [1995], 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *J. R. Stat. Soc. Ser. B* pp. 289–300.

[2] Berger, J. [2006], 'The case for objective Bayesian analysis', *Bayesian Anal.* (3), 385–402.

[3] Berger, J. O. and Bayarri, M. J. [2004], 'The Interplay of Bayesian and Frequentist Analysis', *Stat. Sci.* **19**(1), 58–80.

[4] Berger, J. O. and Wolpert, R. L. [1988], *The Likelihood Principle*.

[5] Dempster, A. P., Laird, N. M. and Rubin, D. B. [1977], 'Maximum likelihood from incomplete data via the EM algorithm', *J. R. Stat. Soc. Ser. B* **39**(1), 1–38.

[6] Deng, A. and Hu, V. [2015], Diluted Treatment Effect Estimation for Trigger Analysis in Online Controlled Experiments, *in* 'Proc. 8th ACM Int. Conf. Web search data Min.'.

[7] Deng, A., Li, T. and Guo, Y. [2014], Statistical Inference in Two-stage Online Controlled Experiments with Treatment Selection and Validation, *in* 'Proc. 23rd Int. Conf. World Wide Web', WWW '14, pp. 609–618.

[8] Deng, A., Xu, Y., Kohavi, R. and Walker, T. [2013], Improving the sensitivity of online controlled experiments by utilizing pre-experiment data, *in* 'Proc. 6th ACM Int. Conf. Web search data Min.', ACM, pp. 123–132.

[9] Efron, B. [1986], 'Why isn't everyone a Bayesian?', *Am. Stat.* **40**(1), 1–5.

[10] Efron, B. [2010], *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Vol. 1, Cambridge University Press.

[11] Efron, B. [2011], 'Tweedie's formula and selection bias', *J. Am. Stat. Assoc.* **106**(496), 1602–1614.

[12] Efron, B. [2013*a*], 'A 250-year argument: belief, behavior, and the bootstrap', *Bull. Am. Math. Soc.* **50**(1), 129–146.

[13] Efron, B. [2013*b*], Empirical Bayes modeling, computation, and accuracy, Technical report.

[14] Efron, B. [2014], 'Frequentist accuracy of Bayesian estimates', *J. R. Stat. Soc. Ser. B* .

[15] Gelman, A. and Hill, J. [2006], *Data analysis using regression and multilevel/hierarchical models*, Cambridge University Press.

[16] Johnson, V. E. [2013], 'Revised standards for statistical evidence', *Proc. Natl. Acad. Sci.* **110**(48), 19313–19317.

[17] Kass, R. and Raftery, A. [1995], 'Bayes factors', *J. Am. Stat. Assoc.* **90**(430), 773–795.

[18] Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T. and Xu, Y. [2012], 'Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained', *Proc. 18th Conf. Knowl. Discov. Data Min.* .

[19] Kohavi, R., Deng, A., Frasca, B., Xu, Y., Walker, T. and Pohlmann, N. [2013], 'Online Controlled Experiments at Large Scale', *Proc. 19th Conf. Knowl. Discov. Data Min.* .

[20] Kohavi, R., Deng, A., Longbotham, R. and Xu, Y. [2014], Seven rules of thumb for web site experimenters, *in* 'Proc. 20th Conf. Knowl. Discov. Data Min.', KDD '14, New York, USA, pp. 1857–1866.

[21] Kohavi, R., Longbotham, R., Sommerfield, D. and Henne, R. M. [2009], 'Controlled Experiments on the Web: survey and practical guide', *Data Min. Knowl. Discov.* **18**, 140–181.

[22] Kruschke, J. K. [2013], 'Bayesian estimation supersedes the t test.', *J. Exp. Psychol. Gen.* **142**(2), 573.

[23] Masson, M. E. J. [2011], 'A tutorial on a practical Bayesian alternative to null-hypothesis significance testing.', *Behav. Res. Methods* **43**(3), 679–90.

[24] Muller, P., Parmigiani, G. and Rice, K. [2006], FDR and Bayesian multiple comparisons rules, *in* '8th World Meet. Bayesian Stat.', Vol. 0.

[25] Murphy, K. P. [2012], *Machine learning: a probabilistic perspective*, MIT press.

[26] Park, T. and Casella, G. [2008], 'The bayesian lasso', *J. Am. Stat. Assoc.* **103**(482), 681–686.

[27] Rouder, J. N. [2014], 'Optional stopping: no problem for Bayesians.', *Psychon. Bull. Rev.* **21**(March), 301–8.

[28] Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. and Iverson, G. [2009], 'Bayesian t tests for accepting and rejecting the null hypothesis.', *Psychon. Bull. Rev.* **16**(2), 225–37.

[29] Scott, J. and Berger, J. [2006], 'An exploration of aspects of Bayesian multiple testing', *J. Stat. Plan. Inference* .

[30] Sellke, T., Bayarri, M. and Berger, J. [2001], 'Calibration of ρ values for testing precise null hypotheses', *Am. Stat.* **55**(1), 62–71.

[31] Senn, S. [2008], 'A note concerning a selection "paradox" of Dawid's', *Am. Stat. Assoc.* **62**(3), 206–210.

[32] Student [1908], 'The probable error of a mean', *Biometrika* **6**, 1–25.

[33] Tang, D., Agarwal, A., O'Brien, D. and Meyer, M. [2010], 'Overlapping Experiment Infrastructure: More, Better, Faster Experimentation', *Proc. 16th Conf. Knowl. Discov. Data Min.* .