Scalable Heterogeneity Detection in Online Experiments

Hammaad Adam[1,2], Merlin Heidemanns[2], Doug Hains[2], James McQueen[2]
[1]MIT, [2]Amazon.com

Online sites typically evaluate the impact of new product features on customer behavior using online controlled experiments (or A/B tests). For many business applications, it is important to detect heterogeneity in these experiments [1], as new features often have a differential impact by customer segment, product group, and other variables. Understanding heterogeneity can provide key insights into causal pathways, enable differential launches to subgroups of customers and products, and motivate future experiments.

Recent methods have focused on estimating heterogeneous impacts [1], but they are difficult to apply to online experiments for two reasons. First, existing methods scale poorly in industry settings. With hundreds of thousands of experiments and billions of observations per year, many existing methods are impractical due to their computational speed and need for supervision. Second, existing methods typically detect heterogeneity based on customer features. However, heterogeneity also exists on other axes. For example, a new online retail feature may increase e-book revenue, but reduce revenue from physical books, which reflects heterogeneity in a dimension of the outcome (i.e., revenue broken out by product group). Current approaches for such heterogeneity (e.g., fixed effects) scale poorly with large numbers of customers and categories. There is thus a need for a scalable method that can detect different types of heterogeneity in online experiments.

A Naive, Scalable Approach Consider a naive but scalable approach to detect heterogeneity: use a simple model (e.g., univariate linear regression) to separately estimate the treatment effect in each category (e.g., customer segment, outcome dimension), then flag categories with the strongest evidence of positive or negative impact (e.g., the smallest p-values). This approach is intuitive, as it identifies the categories that are "most impacted" in an experiment. It is also scalable and applicable to any type of heterogeneity, as it just requires fitting a simple model separately on a few pre-defined categories.

The naive approach has two key flaws. First, it produces many false discoveries (i.e., flags categories that are unaffected by the treatment), as it simultaneously evaluates several hypotheses without accounting for multiplicity. Second, it greatly overestimates the true effects in correctly flagged categories. This exaggeration results from the winner's curse [2]: if two categories have the same true effect, the dimension value in which the estimated effect is larger due to random chance is more likely to be flagged. Multiple testing corrections can limit false discoveries, but do not address impact exaggeration, and are overly conservative in low signal, high noise settings (typical in most A/B tests).

Hierarchical Bayes for Improved Scalable Estimation We propose a hierarchical Bayes approach that addresses the flaws of the naive approach while maintaining its benefits. The main benefits of the naive approach are that it is scalable and broadly applicable. Its limitation lies in ignoring the sampling variance of the treatment effects across multiple categories. Our proposed approach addresses this flaw by modeling the distribution of treatment effects across categories. It shrinks the estimates towards a common mean, reducing false discoveries and effect size exaggeration.

Again, we start by separately estimating the effect in each category. We then define a generative hierarchical model of these estimates and infer the posterior distribution of the underlying true effects conditional on the observed estimates. Let θ_j denote the true effect in category j, $\hat{\theta}_j$ denote its estimate, and $\hat{\sigma}_j^2$ denote the estimated sampling variance of $\hat{\theta}_j$ (which is treated as known). We define the following generative process:

$$\mu \sim g_\mu, \quad \tau \sim g_\tau$$
$$\theta_j | \mu, \tau^2 \sim \mathcal{N}(\mu, \tau^2), \quad \hat{\theta}_j | \theta_j \sim \mathcal{N}(\theta_j, \hat{\sigma}_j^2).$$

In this framework, μ is the common mean that the estimates of θ_j are shrunk towards, while τ^2 controls the amount of shrinkage (the larger the τ^2, the further away from μ the posterior estimates are allowed to be). We define appropriate priors g_μ and g_τ for μ and τ, and obtain a posterior distribution for the true treatment effects θ_j using Markov Chain Monte Carlo (MCMC). This posterior estimation scales well in the number of categories (fitting 10k+ categories within minutes). We then flag categories for which posterior probability of $\theta_j > 0$ is greater than γ (i.e., evidence of positive impact) or smaller than $1 - \gamma$ (i.e., evidence of negative impact), where $\gamma \in (0.5, 1)$ is a user-specified threshold.

Evaluation Here, we demonstrate the performance of our hierarchical approach in semi-synthetic simulations. We consider a common task at Amazon: finding the product groups whose revenue is most impacted in an experiment. We use data from the control group of a real A/B test with 115 product groups and simulate the true effects and treatment group data. We consider several distributions for the true effect sizes, including normal, fat-tailed (using a student-t_3), sparse (where all but five effects are zero), and dependent (product group effects have arbitrary correlation sampled from an LKJ distribution). We compare the performance of our hierarchical method to the naive approach described above—both with and without a multiple testing correction—in Fig. 1. The hierarchical approach uses fixed $\mu = 0$, a $\mathcal{N}^+(0, 1)$ prior for τ, and a threshold of $\gamma = 0.66$.

As expected, the naive approach is prone to false discoveries, often identifies effects with the incorrect sign (inflated Type S error), and greatly exaggerates the true effects (exaggeration ratio $>>$ 1). While a multiple testing correction can reduce false discoveries and Type S errors, it greatly worsens effect exaggeration and has very low statistical power (high Type II error). In constrast, the hierarchical approach reduces the exaggeration ratio and Type S error rate, while significantly increasing power. Notably, the hierarchical approach is robust to common violations of the assumptions of its generative model (e.g., independence of true effects across categories). One limitation is that the hierarchical approach is prone to false discoveries in the sparse setting; however, these false discoveries are less likely to lead to incorrect business decisions, as the estimated effects are shrunk close to 0 (mean absolute value of false discovery $< 10^{-3}$, vs. 0.74 for the naive approach).

Practical Application We showcase a practical application of our hierarchical approach. We consider a series of experiments that tested the same feature in different contexts (e.g., desktop, mobile), and estimate impacts by customer segment. Our estimates (Tab. 1) suggest that customer segments 2 and 4 were positively impacted by the feature in almost all contexts. Meanwhile, customer segments 1 and 3 were more lukewarm, showing heterogeneous behaviour across contexts. These insights were invaluable to the product team in understanding the feature's business impact and designing future experiments.

Discussion We discuss three key points. First, the choice of prior for μ and τ should be driven by the specifics of the domain. We experimented with horseshoe priors for τ and mixture priors for μ, but found that these showed no improvement over the chosen normal priors. Second, the choice of threshold γ reflects the decision tradeoffs of the experimenter.

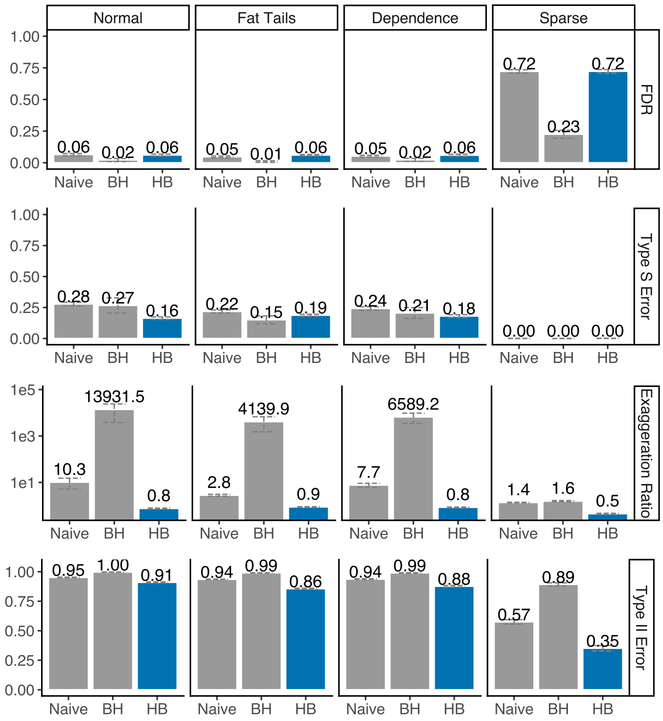

Figure 1: Performance of the hierarchical Bayes ("HB") approach compared to the naive approach ("Naive") and the Benjamini-Hochberg procedure ("BH"). We display mean results across 500 simulated experiments. Error bars denote 95% confidence intervals of the mean.

While thresholding based on posterior probabilities is not decision-theory optimal, it is practical and easily tunable. We found that setting $\gamma = 0.66$ produced valuable insights in exploratory situations where false negatives were less tolerable than false positives. If instead false positives were more costly, a higher threshold (e.g., $\gamma = 0.95$) would be more appropriate. Finally, our approach detects heterogeneity based on pre-defined categories. While this is less flexible than causal estimation methods that detect heterogeneity based on customer features, it is often impractical to operationalize decisions based on arbitrary groups of customers. For example, launching to an arbitrary group requires storing and propagating customer identifiers across all production systems and updating the group with new customers as they arrive. This is often infeasible, especially if identifiers in an experiment are anonymized to maintain privacy. Thus, determining heterogeneity based on pre-defined customer segments (or other categories) leads to more actionable inference.

Experiment	Overall	Segment 1	Segment 2	Segment 3	Segment 4
#1	0.04	0.04	0.08	-0.03	0.12
#2	0.08	0.06	0.12	0.06	0.12
#3	0	0.15	-0.08	0.02	-0.03
#4	0.01	0.01	0.05	-0.11	0.04
#5	0.01	-0.03	-0.01	0.02	-0.01
#6	0.01	0.02	0.11	0.04	0.16

Table 1: Estimates of effect by customer segment of a new feature on sales in six contexts. Colors indicate likely negative (red), indeterminate (gray), and likely positive (green) impacts.

[1] Nicholas Larsen, Jonathan Stallrich, Srijan Sengupta, Alex Deng, Ron Kohavi, and Nathaniel T Stevens. Statistical challenges in online controlled experiments: A review of a/b testing methodology. *The American Statistician*, 78(2):135–149, 2024.

[2] Erik van Zwet and Andrew Gelman. A proposal for informative default priors scaled by the standard error of estimates. *The American Statistician*, 76(1):1–9, 2022.