Estimating the True Effect Size Distribution with SIMEX

Zhaoqi Li[1], Daniel Ting[2], Ilya Gorbachev[2], Ehsan Emamjomeh-Zadeh[2], and Houssam Nassif[2]

[1]Stanford , zli9@stanford.edu
[2]Meta , {dting, ilyagorbachev, ehsanez, houssamn}@meta.com

1 Introduction

Large scale online experimentation inherently runs many experiments. Unfortunately, this also generates a multiple testing problem and results in overstated gains that cannot be reproduced [6]. For experimenters who see their treatment underperform expectations after launch, this can lead to mistrust in the system. Furthermore, experimentation platforms themselves must make a number of policy decisions that depend on their ability to predict the performance of a policy. For example, they may wish to choose a lower significance level for hypothesis tests if the additional product launches yield higher overall gains that exceed the cost of increased false positive rates or Type I error.

While these issues can be addressed, for example by using Bayesian methods to adjust overstated gains or by evaluating decision processes under a generative model, these methods require estimating the distribution of true effect sizes from noisy estimated effect sizes. This paper's main contribution is a novel, non-parametric method to estimate this based on the ideas of SIMulation-EXtrapolation (SIMEX) [3].

Existing work in online experimentation [2, 8] examines the False Discovery Rate of experiments. This problem is also not restricted to online experimentation but is also encountered in the "reproducibility crisis" in academia where efforts like [7, 1, 10] estimate additional properties of a corpus of experiments, such as the power of tests for detecting true effect sizes or the probability that experiments can be replicated.

This problem of estimating the unobserved distribution of true effects is related to the deconvolution problem. In the deconvolution problem, there are independent draws from a density f and additive measurement error from g, so that the observations are drawn from the convolution $h = f * g$. One wishes to estimate the deconvolved density f. Several methods have been developed for this problem, particularly in a line of work using deconvolution kernel density estimators [9, 4], empirical Bayes normal means models (EBNM) [5, 11], or simply normal mixture models. However, estimating effect size distributions in ABTest differ from most deconvolution problems in two major ways. First, it allows for homoskedastic errors. Second, the signal-to-noise ratio is extremely high. The noise distribution is expected to have greater spread than the effect size distribution so that kernel methods are inappropriate for the problem.

More formally, suppose that n experiments are run where the unobserved true average treatment effect is denoted by X_i for experiment i. Further, suppose that the observed ATE estimate $\widehat{X}_i$ is unbiased with known variance σ_i^2. Let $\mathbb{F}_0$ denote the unobserved empirical distribution of the true effects X_i and $\mathbb{F}_{\text{vec}\sigma}$ denote the empirical distribution of the estimated lifts $\widehat{X}_i$. Our goal is to estimate $\mathbb{F}_0$. We can also consider the generative process where X_i is drawn i.i.d. from a distribution

F_0. That is

$$X_i \sim F_0 \tag{1}$$

$$\widehat{X}_i | X_i \sim Normal(X_i, \sigma_i^2). \tag{2}$$

Our goal then is to estimate the distribution F_0.

2 Methodology

Our goal is to learn the CDF of the underlying empirical effect distribution $\mathbb{F}_0$. However, the true measurements X_i of the effect size are never observed. Instead, we observe noisy effect size measurements $X_i + \epsilon_i \sim Normal(X_i, \sigma_i^2)$ and their biased empirical CDF $\mathbb{F}_1$. We then wish to undo the effect of the noise process that takes $\mathbb{F}_0 \to \mathbb{F}_1$.

We first describe the general idea behind bias correction using SIMEX. SIMEX estimates the bias induced by the noise process by adding even more noise. Suppose we were given data $\mathcal{X}$ without measurement error, with some estimator or function $\phi(\mathbf{X})$ of interest. We wish to obtain a good estimator even when the observations $\mathcal{X} + \mathcal{E}$ contain measurement error $\mathcal{E}$. By adding even more noise, we can compute a random function $\theta(c) \stackrel{d}{=} \phi(\mathcal{X} + c \cdot \mathcal{E})$ for any $c \geq 1$. This constitutes the SIMulation component.

It is easy to see that the desired estimate is $\theta(0) = \phi(\mathcal{X})$. Thus, our goal is to Extrapolate the function $\theta(c)$ for $c \geq 1$ to the desired estimate $\theta(0)$. By fitting a parametric regression function $\widehat{\theta}$ on the simulated values $\theta(c)$ for $c > 1$, we can extrapolate to get our final estimate $\widehat{\theta}(0)$. In summary, the general SIMEX procedure is:

1. Simulate noise to compute $\theta(c) = \phi(\mathcal{X} + c\mathcal{E}')$ for $c \geq 1$.
2. Fit a smooth function to $\theta(c)$, $c \geq 1$.
3. Extrapolate to $\widehat{\theta}(0)$.

For our application in effect size estimation, we choose the function of interest ϕ_q to be the q^{th} quantile. We estimate a grid of quantiles to obtain an estimate of the inverse CDF $\mathbb{F}^{-1}$ as opposed to directly trying to estimate the CDF. We found, however, that the typical quadratic regression used in SIMEX was poor at extrapolation. Furthermore, independently estimating each $\phi_q(\mathcal{X})$ would sometimes result in the inverse CDF being non-monotone.

This work introduces two improvements to address these drawbacks. First, our choice to estimate quantiles rather than the CDF allows us to construct a basis that yields both: 1) consistent estimates of the effect size distribution if it belongs to a given parametric family, and 2) good empirical estimates even when it is not in the parametric family. Second, we extend the isotonic regression to ensure all extrapolated estimates $\widehat{F}_c^{-1}$ are monotonic for all c.

3 Experiments

We provide an experiment to illustrate our method. We consider the simple case where the true effect distribution $\mathbb{F}_0 = N(0, 1)$, and measurement error is also $N(0, 1)$. Figure 1 show in solid lines the evolution of each quantile $\theta(c)$ as more noise c is added, while dotted lines show the estimated quantiles as noise is removed through extrapolation. Figure 2 shows that we appropriately tighten the observed distribution of estimated lifts and recover the true effect distribution. We also compare it to EBNM [11] which applies a parametric model, and our non-parametric method is nearly as good. While not shown here, our method also works for other effect size distributions. It also demonstrates some robustness properties as the quantiles underlying the method inherently enjoy some level of robustness.

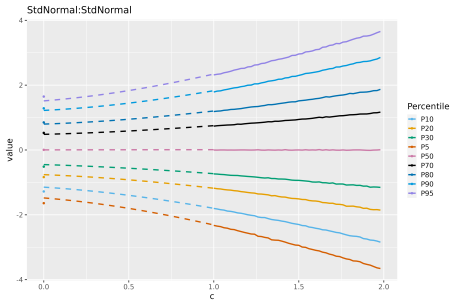

Figure 1: SIMEX curves for different percentiles

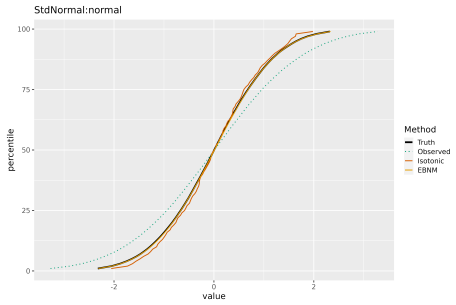

Figure 2: Estimated CDF for different methods

References

[1] František Bartoš and Ulrich Schimmack. Z-curve 2.0: Estimating replication rates and discovery rates. *Meta-Psychology*, 6, 2022.

[2] Ron Berman and Christophe Van den Bulte. False discovery in a/b testing. *Management Science*, 68(9):6762–6782, 2022.

[3] John R Cook and Leonard A Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association*, 89(428):1314–1328, 1994.

[4] Aurore Delaigle. Deconvolution kernel density estimation. In *Handbook of Measurement Error Models*, pages 185–220. Chapman and Hall/CRC, 2021.

[5] Bradley Efron. Empirical bayes deconvolution estimates. *Biometrika*, 103(1):1–20, 2016.

[6] Tanner Fiez, Houssam Nassif, Yu-Cheng Chen, Sergio Gamez, and Lalit Jain. Best of three worlds: Adaptive experimentation for digital marketing in practice. In *The Web Conference (WWW)*, pages 3586–3597, 2024.

[7] Leah R Jager and Jeffrey T Leek. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, 15(1):1–12, 2014.

[8] Ron Kohavi and Nanyu Chen. False positives in a/b tests. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5240–5250, 2024.

[9] Leonard A Stefanski and Raymond J Carroll. Deconvolving kernel density estimators. *Statistics*, 21(2):169–184, 1990.

[10] Erik van Zwet, Andrew Gelman, Sander Greenland, Guido Imbens, Simon Schwab, and Steven N Goodman. A new look at p values for randomized clinical trials. *NEJM evidence*, 3(1):EVIDoa2300003, 2023.

[11] Jason Willwerscheid, Peter Carbonetto, and Matthew Stephens. ebnm: An r package for solving the empirical bayes normal means problem using a variety of prior families. *arXiv preprint arXiv:2110.00152*, 2021.