From Augmentation to Decomposition: A New Look at CUPED in 2023

Alex Deng alex.deng@airbnb.com
Luke Hagar lmhagar@uwaterloo.ca
Nathaniel Stevens nstevens@uwaterloo.ca
Tatiana Xifara tatiana.xifara@airbnb.com
Lo-Hua Yuan lohua.yuan@airbnb.com
Amit Gandhi amit.gandhi@airbnb.com

1) Motivation

Statistical Power continues to be a main challenge and dominating factor in online experimentation quality and velocity

- Variance Reduction techniques like CUPED are widely used, but often in a covariate regression form which limits its applicability and impact
- A new look at CUPED:
  - **CUPED is an Mean-Zero Augmentation Method**
  - The source of Mean-Zero Augmentation can be beyond pre-experiment data, and can consider more than just simple metric differences
  - Mean-Zero Augmentations encode prior knowledge of noises vs. signal

Here we consider Metric Decomposition as an extension of CUPED to **Almost Mean-Zero Augmentation**

- Often times domain knowledge can lead to the decomposition of a metric

$$M = M_1 + M_2,\ \Delta(M) = \Delta(M_1) + \Delta(M_2)$$

where one component captures more impact from an intervention than the other component, while the other component can contribute a significant amount of noise to the metric

- How can we leverage these kinds of decompositions to improve our estimation of treatment effect for M? Can we use CUPED estimators like $\Delta^* = \Delta_2 + \theta\Delta_1$?

2) CUPED as Mean-Zero Augmentation

First Insight: Augmentation with Mean Zero Term

For any estimator $\hat{\delta}$ of δ, we can define a new estimator

$$\hat{\delta}^* = \hat{\delta} + \hat{\delta}_0$$

This augmented estimator has the same mean as the original estimator as long as the augmented term has mean zero. This term can be constructed from any functional form of observations, and does not need to be a difference of the same metric from two treatment variants

Second Insight: Variance Reduction is Guaranteed

Any mean-zero augmentation term yields a whole family

$$\hat{\delta}^*(\theta) = \hat{\delta} + \theta\hat{\delta}_0$$

The optimal θ that minimizes variance is:

$$\theta := -\frac{\mathrm{Cov}\left[\hat{\delta}_0, \hat{\delta}\right]}{\mathrm{Var}\left[\hat{\delta}_0\right]}$$

Reduction proportional to correlation

with variance $\mathrm{Var}\left[\hat{\delta}^*(\theta)\right] = \mathrm{Var}\left[\hat{\delta}\right] \times \left(1 - \frac{\mathrm{Cov}\left[\hat{\delta}_0, \hat{\delta}\right]^2}{\mathrm{Var}\left[\hat{\delta}\right]\mathrm{Var}\left[\hat{\delta}_0\right]}\right) = \mathrm{Var}\left[\hat{\delta}\right] \times \left(1 - \mathrm{Corr}\left[\hat{\delta}, \hat{\delta}_0\right]^2\right)$

Third Insight: Existence of Mean-Zero Augmentation

For controlled experiments, mean-zero augmentation always exists. For any metric M computed from observations X in the pre-experimentation period. The following is a mean-zero augmentation (i.e. $M(X) = \overline{X}$)

$$\Delta_{pre}(M) = M(\mathbf{X}_{pre})_t - M(\mathbf{X}_{pre})_c$$

3) Augmentation View is Better

When M is a simple average metric,

$$\hat{\delta} = \overline{Y}_t - \overline{Y}_c, \hat{\delta}_0 = \overline{X}_t - \overline{X}_c$$
$$\hat{\delta}^* = \hat{\delta} - \theta \cdot \hat{\delta}_0 = \overline{Y - \theta X}_t - \overline{Y - \theta X}_c$$

- CUPED is equivalent to first computing the residuals of a regression and then computing their difference in means.
- This interpretation is limited. CUPED as augmentation applies beyond simple average metric.

Flexible Metric Form (derived metrics)

The estimator is the difference in metric values, but the metric can be a **derived** metric as a function of other averages

Example 1: Ratio metric

$$\Delta = M_t - M_c = \frac{\overline{R}_t}{\overline{S}_t} - \frac{\overline{R}_c}{\overline{S}_c} \qquad \Delta_0 = M_t^{pre} - M_c^{pre} = \frac{\overline{R}_t^{pre}}{\overline{S}_t^{pre}} - \frac{\overline{R}_c^{pre}}{\overline{S}_c^{pre}}$$

CUPED $\Delta + \theta\Delta_0$ no longer has a residualized form

Example 2: Percentile metric

Flexible Augmentation Form

- Augmentation does not need to be from pre-experiment data, and it doesn't need to be based on a difference of metric values

Example: One-sided triggering Some features require opt in, and only treatment group users can opt in. We can construct mean-zero augmentation by **matching** treatment not opt in users with control users using **observational causal inference** techniques! (Deng et. al. 2023)

4) Metric Decomposition: From Mean Zero to Approximate Mean Zero

CUPED variance reduction is determined by correlation between the augmentation term and the original estimator. To yield high correlation, it is best that the original estimator and augmentation share a common component!

If a metric can be decomposed into two components

$$M = M_1 + M_2,\ \Delta(M) = \Delta(M_1) + \Delta(M_2)$$

We can interpret $\Delta(M_2) = \Delta(M) - \Delta(M_1)$

as CUPED with -1 coefficient on the first component Delta as augmentation

- High correlation and variance reduction when first component contributes significantly to the combined metric M
- Delta of first component M_1 is not guaranteed to have mean-zero

Main idea: With domain knowledge of what is the first order effect of the treatment intervention, we can construct such a decomposition so that the first component is affected much less than the second component by the treatment.

In general, the augmentation is $\Delta^* = \Delta(M) - (1-\theta)\Delta_1$

$$\Delta^* = \Delta_2 + \theta\Delta_1$$

What is the **optimal CUPED estimator** in this setting? What is our **objective for optimization** when augmentation is not mean-zero but only **approximately mean-zero**?

5) Bivariate Model and Empirical Bayes

Let $\underset{\sim}{\Delta} = (\Delta_1, \Delta_2)$ be the decomposed vector of $\Delta(M_1)$ and $\Delta(M_2)$

True effect vector $\underset{\sim}{\delta} = (\delta_1, \delta_2)$

Our inference target is the treatment effect on the metric before decomposition $\delta = \delta_1 + \delta_2$

Simple bivariate model:

$$\underset{\sim}{\Delta} \sim \underset{\sim}{\delta} + \underset{\sim}{\varepsilon}$$

By CLT, we assume the noises follow a bivariate Gaussian distribution

$$\underset{\sim}{\varepsilon} \sim N(0, \Sigma)$$

The effect $\underset{\sim}{\delta}$ follows a bivariate distribution with covariance Λ

Both covariance matrices can be estimated from data from a set of historical experiments

6) Bayesian Variance Reduction

What difference does metric decomposition make?

Under the frequentist framework, it doesn't change inference. Because if our inference target is $\boldsymbol{\delta}$ = E($\Delta(M)$), using the normal likelihood, $\Delta(M)$ is already the sufficient statistic so there is no efficiency gain from further decomposition

Decomposition changes Bayesian posterior!

Univariate (No decomposition) $\delta|\Delta_1 + \Delta_2$

Bivariate (decomposition) $\delta|\Delta_1, \Delta_2$

Assuming a normal prior, we can work out the posterior mean and variance for both cases

Univariate: $\mathrm{E}[\delta|\Delta] = A\Delta\,, \quad \mathrm{Var}[\delta|\Delta] = A\sigma^2$

$$A = \frac{\lambda^2}{\lambda^2+\sigma^2},\ \lambda^2 = \Lambda_{11}+\Lambda_{12}+\Lambda_{21}+\Lambda_{22} \text{ and } \sigma^2 = \Sigma_{11}+\Sigma_{12}+\Sigma_{21}+\Sigma_{22}$$

Bivariate: $\mathrm{E}\left[\delta|\underset{\sim}{\Delta}\right] = 1^T \cdot S\underset{\sim}{\Delta}\,, \quad \mathrm{Var}\left[\delta|\underset{\sim}{\Delta}\right] = 1^T \cdot (I-S)\Lambda \cdot 1$

$$S = \Lambda(\Lambda+\Sigma)^{-1}.$$

We showed the bivariate posterior variance is always smaller than the univariate posterior variance, especially when one component is approximately mean-zero and its prior variance is small

7) Frequentist Optimal CUPED

Bayesian posterior mean has shrinkage form $\alpha\Delta_1 + \beta\Delta_2$

Frequentist CUPED only shrinks augmentation: $\Delta^* = \Delta_2 + \theta\Delta_1$

Two slightly different optimization objectives

1) **minimizing prediction error** $\arg\min_\theta E[((\Delta_2 + \theta\Delta_1) - \delta)^2]$

2) **maximizing correlation** $\arg\max_\theta Corr(\Delta_2 + \theta\Delta_1, \delta)$

7) Frequentist Optimal CUPED (cont'd)

Recall Notation: $Cov(\underset{\sim}{\delta}) = \Lambda = \begin{pmatrix}\Lambda_{11} & \Lambda_{12}\\ \Lambda_{21} & \Lambda_{22}\end{pmatrix}, \quad Cov(\underset{\sim}{\epsilon}) = \Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$

Minimizing Prediction error: $\theta = \frac{\Lambda_{11} - \bar{\Sigma}_{12}}{\Lambda_{11}+\Sigma_{11}}$

Extreme case 1: when Λ_{11} close to 0, this reduce to CUPED when augmentation term has mean 0.

Extreme case 2: when Σ_{12} small, this reduce to Bayesian shrinkage for Δ_1

This is a direct extension of CUPED when augmentation is approximately mean-zero

Maximizing Correlation:

$$\theta = \frac{(\Lambda_{12}+\Sigma_{12})(\Lambda_{12}+\Lambda_{22}) - (\Lambda_{11}+\Lambda_{12})(\Lambda_{22}+\Sigma_{22})}{(\Lambda_{12}+\Sigma_{12})(\Lambda_{11}+\Lambda_{12}) - (\Lambda_{12}+\Lambda_{22})(\Lambda_{11}+\Sigma_{11})}.$$

We showed that **this is a rescaled bivariate Bayesian posterior mean!**

i.e. Bivariate Bayesian posterior with decomposition is $\alpha\Delta_1 + \beta\Delta_2$

To only shrink first component, divide by α and let $\theta = \beta/\alpha$

8) Real World Application

Ideas to construct approximate mean-zero component?

1. The first order treatment effect often requires a subject to direct interaction with a feature. Merely exposure to the feature only yields second order effects
2. Model based causal surrogate as proxy to a target metric. The residual of the surrogate metric is approximate mean-zero if causal surrogacy assumption almost holds
3. When testing ML algorithms, counterfactual ML output can be leveraged. We can identify cases where the treatment and control will have similar potential outcomes

Airbnb ranking experiments: we utilize ranking results from both treatment and control ranker and user's click location (map vs. feed) to construct the approximate mean-zero component and decompose booking metric into two parts.

Among 25 real experiments, estimated prior covariance and noise covariance matrix (rescaled for easier comparison):

$$\Lambda = \begin{pmatrix}0.576 & -0.896\\ -0.896 & 4.329\end{pmatrix} \quad \text{and} \quad E[\Sigma] = \begin{pmatrix}4.020 & 0.169\\ 0.169 & 0.811\end{pmatrix}$$

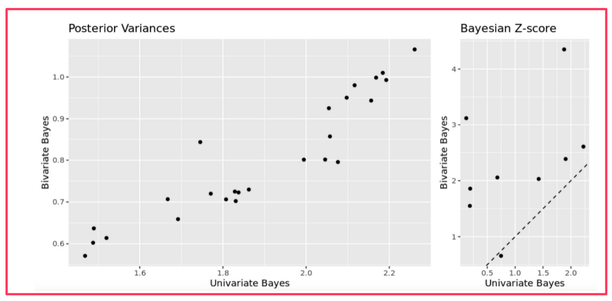
1) We see first component has a much smaller true effect variance than the second component.
2) First component has a much larger noise variance than the second component.
3) Ratio of effect variance to noise variance measures signal noise ratio. First component SNR is only 0.14, second component is 5.34, before decomposition only 0.60

The following table shows the number of stat. sig. results out of 25 experiments. Using second the component alone, using CUPED maximizing correlation or minimizing error all lead to more stat. sig. results than without decomposition (8 vs. 4) . First component had 8% (2/25) stat. sig. not far from 5% if it is an exact mean-zero augmentation

Number of Stat. sig. results

Δ	Δ_1	Δ_2	CUPED_{corr}	CUPED_{err}
4/25	2/25	8/25	8/25	8/25

Bivariate Bayesian posterior variance is reduced by more than 50% compared to univariate Bayesian variance without decomposition. As a result our decomposition method lead to more precise Bayesian credible interval and higher Bayesian z-score



12) Final Remarks

- CUPED augments any treatment effect estimator by a zero-mean component. This view is more flexible and general.
- We extend CUPED to settings where augmentation is only approximately with mean-zero. Specifically, when the metric can be decomposed into two components where one component is mostly noise.
- Metric decomposition always leads to reduced posterior variance
- Optimize for minimizing prediction error or maximizing correlation gives different forms of CUPED type optimal augmentation.