Metric Decomposition in A/B Tests

Alex Deng*
Airbnb
Seattle, WA, USA
alex.deng@airbnb.com

Luke Hagar
University of Waterloo
Waterloo, ON, Canada
lmhagar@uwaterloo.ca

Nathaniel T. Stevens
University of Waterloo
Waterloo, ON, Canada
nstevens@uwaterloo.ca

Tatiana Xifara
Airbnb
San Francisco, CA, USA
tatiana.xifara@airbnb.com

Amit Gandhi†
University of Pennsylvania
Philadelphia, PA, USA
agandhi@upenn.edu

Abstract

More than a decade ago, CUPED (Controlled Experiments Utilizing Pre-Experiment Data) mainstreamed the idea of variance reduction leveraging pre-experiment covariates. Since its introduction, it has been implemented, extended, and modernized by major online experimentation platforms. Despite the wide adoption, it is known by practitioners that the variance reduction rate from CUPED utilizing pre-experimental data varies case by case and has a theoretical limit. In theory, CUPED can be extended to augment a treatment effect estimator utilizing in-experiment data, but practical guidance on how to construct such an augmentation is lacking. In this article, we fill this gap by proposing a new direction for sensitivity improvement via treatment effect augmentation whereby a target metric of interest is decomposed into components with high signal-to-noise disparity. Inference in the context of this decomposition is developed using both frequentist and Bayesian theory. We provide three real world applications demonstrating different flavors of metric decomposition; these applications illustrate the gain in agility metric decomposition yields relative to an un-decomposed analysis.

CCS Concepts

• **Mathematics of computing** → **Probabilistic inference problems**; **Hypothesis testing and confidence interval computation**; *Bayesian computation*; • **Applied computing** → **E-commerce infrastructure**.

Keywords

A/B testing, online experimentation, variance reduction, Bayesian analysis, causal surrogate, counterfactual

ACM Reference Format:

Alex Deng, Luke Hagar, Nathaniel T. Stevens, Tatiana Xifara, and Amit Gandhi. 2024. Metric Decomposition in A/B Tests. In *Proceedings of Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3637528.3671556

*Corresponding author.
†Work completed while employed by Airbnb.

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671556

1 Introduction

Online controlled experiments, also referred to as "A/B tests", are an invaluable tool used by companies to test and evaluate changes to their online products. With respect to some metric(s) of interest, these experiments facilitate causal conclusions about the efficacy of such changes. Large tech companies collectively run tens of thousands of these experiments each year, engaging millions of users [23].

An A/B test typically compares two versions of a product: a new *treatment* version to the existing *control* version. Interest lies in understanding the *treatment effect* δ, which quantifies the potential improvement (with respect to some metric of interest) induced by the treatment relative to the control. Denoting the metric of interest M, the treatment effect δ is commonly estimated using the difference in metric values observed in the treatment and control groups $\Delta(M) := M_T - M_C$. Assuming the users are independent of one another and randomized to the treatment and control groups, this estimator is unbiased for δ. In some contexts, a treatment effect defined on the percent scale is preferred for ease of business communication. This is referred to as *lift*, and is estimated by

$$\Delta\%(M) := \frac{M_T - M_C}{M_C}.$$

In A/B tests, such metrics are often defined as averages $M := \overline{X}$ of some measurement X_i observed on each user $i = 1, 2, \ldots, n$ in the treatment (or control) group. However, ratio and percentile metrics may also be relevant [9, 22, 23]. In this paper, we focus on average and ratio metrics as they account for the appreciable majority of metrics used in practice.

Thus, inference (by way of hypothesis tests and statistical intervals) for δ using $\Delta(M)$ (or $\Delta\%(M)$) is of interest. However, such inference is complicated by the noisiness of these metrics in practice; inference quality hinges critically on the sampling variances

$$\mathrm{Var}[\Delta(M)] \quad \text{and} \quad \mathrm{Var}[\Delta\%(M)].$$

Although sample sizes in online A/B tests are typically very large—often at least thousands up to millions—it is widely documented by practitioners that metrics of interest are highly variable and that hypothesis tests for δ lack statistical power [20]. Consequently, false negatives—when experimenters cannot detect a non-zero treatment effect—are prevalent. Moreover, in the face of statistically significant (abbreviated as stat. sig. henceforth) results, the estimated treatment effect $\Delta(M)$ (or $\Delta\%(M)$) often over exaggerates the true treatment effect δ yielding false discoveries [18, 21].

Therefore, there is great interest in increasing sensitivity of metrics; for a given metric M, determining how to construct an estimator of the treatment effect δ with low bias and small variance remains one of the most critical statistical challenges for A/B testing research [3, 20, 23]. Assuming an unbiased estimator, the method most widely applied in industry to reduce variability is CUPED (Controlled experiments Utilizing Pre-Experimental Data) or its variants and extensions [7, 12, 13]. The general idea with this class of methods is to use in place of M an alternative version of the metric that is augmented by a *second* metric highly correlated with M. Another class of methods recently gaining traction is the use of *surrogate metrics* in place of M. Such surrogates are chosen or designed to be proxies for M with higher sensitivity [15].

In this paper, we propose novel methodology for increasing the sensitivity of metrics and hence treatment effect estimators that represents a new direction on this problem. In particular, we propose decomposing the metric of interest into two or more components in an attempt to isolate those with high signal and low noise from those with low signal and high noise. The paper demonstrates both empirically and theoretically the value of this practice in both frequentist and Bayesian settings.

1.1 Metric Decomposition

Consider an additive decomposition of a metric M as follows $M = M_1 + M_2$. This decomposition implies the following decomposition of the estimator

$$\Delta(M) = \Delta(M_1) + \Delta(M_2)\,, \tag{1}$$

where the true effect to be estimated also has the decomposition $\delta = \delta_1 + \delta_2$. Multiplicative decompositions such as $M = M_1 \times M_2$ may also be of interest. In a treatment vs. control comparison, if we observe percent lifts $\Delta\%(M_1)$ and $\Delta\%(M_2)$, by the multiplicative decomposition, we have

$$\begin{aligned} M_T &= M_{1,T} \times M_{2,T} \\ &= [1+\Delta\%(M_1)]M_{1,C} \times [1+\Delta\%(M_2)]M_{2,C} \\ &= M_C \times [1+\Delta\%(M_1)][1+\Delta\%(M_2)]\,. \end{aligned}$$

Dividing both sides by M_C and expanding the right hand side, we see

$$\frac{M_T}{M_C} - 1 = \Delta\%(M_1) + \Delta\%(M_2) + \Delta\%(M_1)\cdot\Delta\%(M_2)\,.$$

The left hand side is the percent treatment effect $\Delta\%(M)$, and the last term on the right hand side is an ignorable second order term. When both $\Delta\%(M_1)$ and $\Delta\%(M_2)$ are relatively small, which is often the case in A/B tests where even a 10% change is commonly deemed extreme, the following *approximate* additive decomposition is appropriate

$$\Delta\%(M) \approx \Delta\%(M_1) + \Delta\%(M_2). \tag{2}$$

With a unified (though slight abuse of) notation, we let $\delta \approx \delta_1 + \delta_2$ represent the ground truth multiplicative treatment effect. Thus decompositions of $\Delta(M)$ and $\Delta\%(M)$ will both be treated as additive no matter whether the decomposition of M is additive or multiplicative.

Note that the above decompositions assume the metric M decomposes into $k = 2$ components, but context and/or engineered solutions may dictate a decomposition into any number of components, e.g.,

$$M = M_1 + \cdots + M_k \quad \text{or} \quad M = M_1 \times \cdots \times M_k.$$

We address (and develop theory for) this more general case in this paper.

Where does a decomposition come from?
Context may dictate a natural decomposition. If M is a simple average $\overline{X}$, an additive decomposition can come from breaking each observation X_i into two (or more) parts. Similarly, when M is a ratio metric $\overline{X}/\overline{Y}$, an additive decomposition can be constructed from a decomposition of the numerator $\overline{X}$. Multiplicative decompositions such as $\overline{X} = \frac{\overline{X}}{\overline{Y}} \times \overline{Y}$ also occur naturally. For example, multiplicative metric chaining is common in conditional funnels; if a conversion funnel involves multiple steps, then a conversion rate $\overline{Y}$ at an intermediate step can be used to decompose the overall conversion rate $\overline{X}$ multiplicatively. Similarly, common revenue metrics such as revenue per user can be decomposed into revenue per purchase, and purchases per user.

More generally, both additive and multiplicative decompositions can be *engineered* by defining one of the components and then taking the second component to be the additive or multiplicative complement. Specifically, let M_1 be any arbitrary metric, we can define M_2 as $M - M_1$ (in the additive case), or M/M_1 (in the multiplicative case). With this construction, we can always get a *synthetic* metric decomposition. In Section 3, we illustrate real-world examples of both contextual and engineered decompositions.

Why is metric decomposition useful?
In Sections 2 and 4 we respectively develop frequentist and Bayesian theory for how to leverage decompositions to improve the sensitivity of treatment effect estimators relative to the standard approach without decomposition. Then, in Section 3, we provide three real-world applications of metric decomposition to illustrate how these methods can be employed in practice. Here, we motivate at a high-level the value of metric decomposition from both the frequentist and Bayesian perspectives.

In the frequentist setting, we propose defining a new treatment effect estimator as a function of the components. Illustrating the basic idea with $k = 2$ components, we have

$$\Delta^*(\theta) := \Delta_1 + \theta \cdot \Delta_2 \tag{3}$$

where Δ_1 and Δ_2 are suppressed notation for $\Delta(M_1)$ and $\Delta(M_2)$. Clearly, the original estimator from (1) arises as a special case when $\theta = 1$, but the formulation in (3) allows for optimization of different objectives with respect to θ. Such objectives may include variance reduction, mean squared error reduction, or power boosting. As we demonstrate in Section 2.2, the proposed framework is flexible to a variety of different objectives.

Keen readers familiar with variance reduction and CUPED [12] will recognize this form of regression adjustment. When the component Δ_2 has no treatment effect, i.e., $\delta_2 = 0$, we know $\delta = \delta_1$. Instead of using $\Delta = \Delta_1 + \Delta_2$, we can directly use Δ_1 as an estimator if it has a smaller variance than Δ. Or, more generally, we can find θ that minimizes variance in the family of (3). However, in general, beyond using pre-experiment data as suggested by CUPED, it is hard to construct a component Δ_2 with theoretically 0 treatment effect. Nevertheless, as we demonstrate in this paper, it is commonly possible to define a decomposition in which the components have drastically different signal-to-noise ratios (SNRs), with one component capturing the majority of the treatment effect and the other component(s) capturing much less treatment effect and a large proportion of noise. With metric decomposition, we can exploit this kind of *SNR disparity*. In this way, the estimator in (3) based on metric decomposition can

be seen as a generalization of CUPED, where rather than adjusting by a null-effect term (i.e., mean-zero augmentation) [13], we adjust by an *almost* null-effect term. We refer to the adjustment made with estimators in the form of (3) as *approximately null augmentation*, or ANA. This perspective will be formalized in Section 2.

From a Bayesian perspective, inference for δ is carried out via posterior analyses. Of interest here are the two posterior distributions

$$p(\delta|\Delta) \text{ and } p(\delta|\Delta_1,\Delta_2),$$

where the first would be used in a standard analysis and the second exploits the decomposition. As we formalize in Section 4, we can expect the posterior distribution $p(\delta|\Delta_1,\Delta_2)$ to have smaller posterior variance than $p(\delta|\Delta)$. Moreover, certain prior information can also lead to $p(\delta|\Delta_1,\Delta_2)$ being concentrated more closely around δ than $p(\delta|\Delta)$. Thus, from a Bayesian perspective, sensitivity and hence the quality of inference can also be improved by metric decomposition.

1.2 Setup and Notation

We assume the target of inference is the treatment effect δ, which quantifies the additive (or percent) difference between treatment and control with respect to some metric M. We further assume M decomposes into a sum (or product) of components $M_1,\ldots,M_k$. In either case, as discussed in Section 1.1, we assume that $\Delta(M) = \Delta_1(M)+\cdots+\Delta_k(M)$ estimates the treatment effect δ which similarly decomposes: $\delta = \delta_1 + \cdots + \delta_k$. Although in this paper we will illustrate metric decomposition for the basic $k=2$ component version, we develop theory for the general $k > 2$ component case as well. Throughout we'll use the vector notation $\boldsymbol{\Delta} = (\Delta_1,\ldots,\Delta_k)$ to represent observed treatment effects and $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_k)$ to represent true treatment effects. We adopt the following random effect model

$$\boldsymbol{\Delta} = \boldsymbol{\delta} + \boldsymbol{\varepsilon}\,, \tag{4}$$

where $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ are both random vectors which are assumed to be independent of one another. This model is meant to characterize the variation in observed treatment effects across a population of A/B tests (e.g., across all the A/B tests run by a given organization). The random vector $\boldsymbol{\delta}$ reflects variation in true treatment effects across these experiments and the random vector $\boldsymbol{\varepsilon}$ reflects noise in treatment effect estimation. For large scale A/B tests, it is common to exploit the central limit theorem and assume that $\boldsymbol{\varepsilon} \dot{\sim} \mathcal{N}(\mathbf{0},\Sigma)$. We further follow industry convention and assume the covariance matrix Σ is fixed and known, where the "known" values are found using sample variances and covariances based on past experiments. For a given experiment, this covariance matrix is a function of sample size n though we suppress notation and do not notate this dependence explicitly.

We also posit that $\boldsymbol{\delta}$ follows a distribution with mean $\mathrm{E}[\boldsymbol{\delta}] = \mathbf{0}$ and covariance matrix $\mathrm{Var}[\boldsymbol{\delta}] = \boldsymbol{\Lambda}$. Note that unlike $\boldsymbol{\varepsilon}$, which follows a normal distribution due to the central limit theorem, we do not in general assume $\boldsymbol{\delta}$ follows a normal distribution[1]. The zero-mean assumption reflects the reality that across an organization's population of A/B tests, results will be positive, negative, null, and likely null on average. We remark that $\boldsymbol{\Lambda}$ can be estimated empirically from data. For example, suppose we observe N equal sample-sized experiment results each with the observed vector $\boldsymbol{\Delta}_s$, $s=1,\ldots,N$. By the independence of $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$, the covariance of the observed $\boldsymbol{\Delta}$ has a trivial decomposition

$$\mathrm{Var}[\boldsymbol{\Delta}] = \mathrm{Var}[\boldsymbol{\delta}] + \mathrm{Var}[\boldsymbol{\varepsilon}] = \boldsymbol{\Lambda} + \Sigma\,.$$

[1] A normality assumption for $\boldsymbol{\delta}$ is however made in Section 4 when we take a Bayesian view of the problem.

This leads to a sample estimate of $\boldsymbol{\Lambda}$ defined as the difference between the sample covariance matrix of $\boldsymbol{\Delta}$ and the noise covariance matrix Σ. When sample sizes for the set of experiments are different, we can instead use a sample average of Σ_s, $s=1,\ldots,N$. There exist other and more robust ways to estimate $\boldsymbol{\Lambda}$, but this line of research is orthogonal to the metric decomposition methods we propose here. In this paper we'll simply assume an estimate of $\boldsymbol{\Lambda}$ is already available.

We end this section by emphasizing the symbol δ will be used to describe the true unknown treatment effect in a *single* experiment, but also the random variable representing variation in true treatment effects across a population of experiments. Though we take care to distinguish this, context should dictate which version of δ is being used.

1.3 Contributions

This paper makes the following contributions to the online experimentation and measurement science literature:

(1) A new framework for treatment effect estimation that exploits metric decomposition from both frequentist (see Section 2) and Bayesian (see Section 4) perspectives. We also share the code to implement and reproduce our simulation studies[2].
(2) Real-world applications of this new methodology in three different flavors: (i) engineered decomposition, (ii) natural funnel decomposition, and (iii) adjustment of a surrogate metric. These applications are detailed Section 3.

2 Frequentist View of Metric Decomposition

Here we overview the frequentist motivation for metric decomposition. See Section 4 for an elaboration of the Bayesian motivation for decomposition.

2.1 Approximately Null Augmentation

In Section 1.1 we argued that metric decomposition can exploit disparity in signal-to-noise ratios (SNRs). We define the SNR as

$$\frac{\mathrm{Var}[\delta]}{\mathrm{Var}[\varepsilon]}. \tag{5}$$

Given a decomposition with effect variance $\boldsymbol{\Lambda}$ and noise variance Σ, if one component, say the first component (without loss of generality), has an SNR Λ_{11}/Σ_{11} that is much larger than the other components, the intuition is that Δ_1 is the most useful component for estimating δ. And although the other components provide much less signal, they can still be useful as an (approximately null augmentation) adjustment to Δ_1. This leads us to the following definition.

Definition 2.1 (Approximately Null Augmentation)**.** ANA refers to the family of estimators $\Delta^*(\boldsymbol{c}) := \boldsymbol{c}^\intercal\boldsymbol{\Delta}$ where $\boldsymbol{c}^\intercal\boldsymbol{e}_1 = 1$ for the standard basis vector $\boldsymbol{e}_1 \in \mathbb{R}^k$. Note that when $k=2$, this reduces to (3) with $\boldsymbol{c} = (1,\theta)$.

The theoretical results developed in this section address the following question of theoretical and practical importance. For the purpose of estimating $\delta = \mathbf{1}_k^\intercal\,\boldsymbol{\delta}$ in a manner that leverages the SNR disparity and approximately null augmentation, what vector $\boldsymbol{c}$ should we use? In Section 2.2, we define five potential objectives that may be used to find the optimal augmentation vector $\boldsymbol{c}$, and for each we state the optimal coefficients. In Section 2.3 we explore how the proposed metric decomposition method is related to and different from existing variance reduction methods like CUPED and the use of more sensitive surrogate metrics.

[2] https://github.com/lmhagar/MetricDecomp

2.2 ANA Objectives

Note that for brevity and consistency with the examples in Section 3, we consider the $k = 2$ case here, and hence define the optimal θ for each objective. In Appendix A we provide the corresponding derivations and also consider the more general $k > 2$ case.

Minimizing Mean Squared Error. We consider minimizing the MSE of the ANA estimator: $\mathrm{E}[(\delta-\Delta^*)^2]$. Doing so balances bias and variance for better point estimation of effect size at the organizational level. This is a standard regression objective, except that the response δ is not directly observed. However, because the solution for the regression coefficients involves only the covariance of δ and the regressors $\mathbf{\Delta}$, which can all be estimated, we are still able to compute the regression coefficients. See also [5, 26, 28]. With respect to model (4), the value of θ that minimizes $\mathrm{E}[(\delta-\Delta^*)^2]$ is

$$\theta = \frac{\Lambda_{22}-\Sigma_{12}}{\Lambda_{22}+\Sigma_{22}}. \tag{6}$$

Maximizing Correlation. Tripuraneni et al. [28] in a slightly different context suggest maximizing $\mathrm{Corr}[\delta,\Delta^*]$, the correlation between δ and Δ^*. This criterion is useful from the perspective of treating Δ^* as a surrogate metric not just for estimating δ but also for understanding the direction (i.e., sign) of the effect. With respect to model (4), the value of θ that maximizes $\mathrm{Corr}[\delta,\Delta^*]$ is

$$\theta = \frac{(\Lambda_{12}+\Sigma_{12})(\Lambda_{11}+\Lambda_{12})-(\Lambda_{12}+\Lambda_{22})(\Lambda_{11}+\Sigma_{11})}{(\Lambda_{12}+\Sigma_{12})(\Lambda_{12}+\Lambda_{22})-(\Lambda_{11}+\Lambda_{12})(\Lambda_{22}+\Sigma_{22})}. \tag{7}$$

It is also interesting to point out that the ANA maximizing correlation is just a rescaled version of the posterior mean $\mathrm{E}[\delta|\mathbf{\Delta}]$.

Minimizing Error Variance. Whereas minimizing MSE will inherently address the bias-variance trade-off associated with approximately null augmentation, another sensible objective would be to directly minimize the error variance $\mathrm{Var}[\boldsymbol{c}^\intercal\boldsymbol{\varepsilon}]$. This serves as a *lower bound* for what variance reduction is possible, as it corresponds to the optimal adjustment in CUPED. With respect to model (4), the value of θ that minimizes $\mathrm{Var}[\boldsymbol{c}^\intercal\boldsymbol{\varepsilon}]$ is

$$\theta = \frac{-\Sigma_{12}}{\Sigma_{22}}. \tag{8}$$

Maximizing Expected Squared Z-Score. The test statistic associated with $H_0 : \delta = 0$ in an ANA analysis is the following Z-score: $\Delta^*/\sqrt{\mathrm{Var}[\varepsilon_1+\theta\varepsilon_2]}$. To increase the sensitivity of this test we may seek to find the augmentation that maximizes the expected magnitude of this test statistic. We operationalize this by maximizing the expected square of this test statistic, which based on model (4) is $\mathrm{Var}[\Delta^*]/\mathrm{Var}[\varepsilon_1+\theta\varepsilon_2]$. The optimal value of θ for this objective is

$$\theta = \frac{-b-\sqrt{b^2-4ac}}{2a} \tag{9}$$

with $a=\Lambda_{22}\Sigma_{12}-\Lambda_{12}\Sigma_{22}$, $b=\Lambda_{22}\Sigma_{11}-\Lambda_{11}\Sigma_{22}$, $c=\Lambda_{12}\Sigma_{11}-\Lambda_{11}\Sigma_{12}$.

Maximizing Power. While maximizing the expected Z-score seeks to increase sensitivity when testing $H_0 : \delta = 0$, this may be achieved more directly by maximizing power. Whereas the previous objective marginalizes over the distribution of possible δ values, we may seek to maximize the Z-score for a specifically selected (positive) $\dot{\delta}$ that reflects (for instance) an anticipated effect size of interest. Exploiting the decomposition $\dot{\delta} = \dot{\delta}_1 + \dot{\delta}_2$, the test statistic for this anticipated effect is $(\dot{\delta}_1 + \theta\dot{\delta}_2)/\sqrt{\mathrm{Var}[\varepsilon_1+\theta\varepsilon_2]}$. The value of θ that maximizes this test statistic (and hence power) is

$$\theta = \frac{\dot{\delta}_1\Sigma_{12}-\dot{\delta}_2\Sigma_{11}}{\dot{\delta}_2\Sigma_{12}-\dot{\delta}_1\Sigma_{22}}. \tag{10}$$

Note that rather than specifying a *single* effect of interest $\dot{\delta}$[3], a continuum of $\dot{\delta}$ values could be specified and we could maximize an "integrated" test statistic that aggregates across the plausible $\dot{\delta}$ values. In this case the optimal θ is given by (10) but with $\dot{\delta}_1$ and $\dot{\delta}_2$ replaced by $\bar{\dot{\delta}}_1$ and $\bar{\dot{\delta}}_2$ which denote the average of the $\dot{\delta}_1$ and $\dot{\delta}_2$ values across the continuum of interest.

We acknowledge that maximizing power and the expected squared Z-score will lead to an increase in test rejection, but they may lead to increasingly biased point estimates of the true effect for the undecomposed metric, especially when θ is far away from the region [0,1]. Therefore, we recommend bounding θ within [0,1] when optimizing for power or the expected squared Z-score.

When choosing among objectives, one must recognize that there is no uniformly superior objective; which is appropriate depends on a practitioner's goals. If interest lies in accurately and precisely estimating the treatment effect, minimizing MSE is a sensible objective; a practitioner primarily interested in determining the sign of the effect may seek to maximize correlation; and a practitioner interested in increasing test sensitivity may seek to maximize power. That said, in practice, if we are able to find decompositions with very high SNR disparities, ANA with different objectives will not be materially different. This is illustrated in the first two applications in Section 3.

It's also important to emphasize that all of these objectives are defined with respect to model (4). This means that the optimal augmentations are optimal at the organizational level. This does not necessarily imply that the objectives are satisfied at the individual experiment level. In future work, we plan to use simulation to investigate the extent to which the objectives are/ are not satisfied for *individual* experiments.

2.3 Relation to Exisiting Work

Metric decomposition follows from existing work aimed at increasing metric sensitivity and statistical power. It is closely related to CUPED, in that it augments the treatment effect estimator in the interest of improving sensitivity. Metric decomposition can also be viewed as a more sensitive *surrogate metric* of the original metric of interest. Procedures to find the optimal augmentation $\boldsymbol{c}$ using a set of historical experiment results is also a form of *meta-analysis*, and is related to an empirical Bayesian analysis of experiments. In the subsections below, we describe these connections to existing methodology in more detail.

2.3.1 CUPED The CUPED method [12] was inspired by the method of control variates from stochastic simulation [1, 25]. CUPED is a model-free method that relies only on the key observation that any pre-experiment difference between two randomized groups is pure noise due to randomization and should be 0 in expectation as it estimates a null effect. Deng et al. [13] formulated CUPED as mean-zero augmentation:

$$\Delta^* = \Delta + \theta\cdot\Delta_0\ . \tag{11}$$

This is similar to the two-component ANA in (3), but it differs in that the augmentation term Δ_0 in CUPED is assumed to be exactly zero in expectation. From equation (6) we see that the optimal ANA that minimizes MSE contains the optimal CUPED adjustment

[3] In the applications in Section 3, we take $\dot{\delta}_1$ to be the 95th percentile of $\mathcal{N}(0,\Lambda_{11})$ (calibrating to detect reasonably large effects) and $\dot{\delta}_2 = \Lambda_{12}\dot{\delta}_1/\Lambda_{11}$ (the mean of the conditional distribution of $\delta_2|\delta_1=\dot{\delta}_1$).

$(-\mathrm{Cov}[\varepsilon_1,\varepsilon_2]/\mathrm{Var}[\varepsilon_2])$ as a special case when the ANA is in fact an exact mean zero augmentation (i.e., when $\Lambda_{22}=\mathrm{Var}[\delta_2]=0$). This is the ANA that minimizes error variance in (8). Importantly, the augmentation term in ANA also need not come from pre-experimental data.

ANA is a nontrivial extension and a fundamentally different way to construct augmentations, often with a much greater variance reduction possible. Beyond using pre-experiment period data or relying on triggering conditions [8, 13], there aren't many ways to construct true mean-zero augmentations. It is documented by various sources (e.g., [4, 7]) that the amount of variance reduction elicited by CUPED varies and in many cases can be as little as 10% or less. Recently, it has also been shown that CUPED using pre-experiment data has a variance reduction limit [27]. Greater variance reduction can only be achieved with augmentation terms from in-experiment signals. Metric decomposition stems from the idea of using in-experiment observations to directly construct approximately null components with low SNRs. As we have seen, the theory of ANA gives the optimal adjustment to suit a variety of objectives.

2.3.2 Surrogate Metrics Instead of restricting attention to unbiased estimators for a target metric M, the surrogate metric literature (e.g., [2, 5, 7, 15, 28]) aims to use another metric—which may be an existing metric, a functional combination of a set of metrics, or a model prediction of the target metric—as a proxy. Surrogate metrics can often achieve greater variance reduction and greatly improve experimentation agility when the target metric has low statistical power. However, one drawback is that surrogate metrics are generally biased, with the degree of bias varying case by case. Choosing and evaluating surrogate metrics is an active research area in the A/B testing community [23].

Metric decomposition shares the similar goal of using a potentially biased estimator in pursuit of increased sensitivity. ANA differs from methodologies in the surrogate metric literature, however, because we define explicitly how these potentially biased estimators arise by breaking the target metric into components, instead of choosing from a cohort of existing metrics or linear combinations thereof. Moreover, the metric decomposition approach leads to further improvement for *any* surrogate metric, because the target metric can be decomposed by the surrogate and its residual (additive or multiplicative). In this way, ANA can be applied to further adjust any surrogate metric by its residual. We elaborate on this in more detail, and provide a real example in Section 3.3.

2.3.3 Meta Analysis and Empirical Bayes The way we use a set of historical experiments to aid metric development is a form of *meta-analysis* [11, 17, 28]. The framework of estimating the parameters of the distribution of the treatment effect $\boldsymbol{\delta}$ is also a form of empirical Bayes [6, 10, 14, 16, 24, 29] and multilevel (hierarchical) modeling [19]. But to the authors' knowledge, we are the first to study the implications of replacing an observed metric value with a decomposed vector.

3 Real-world Applications

To apply metric decomposition with approximate null augmentation, we require one or more *approximately null components* (ANCs) (i.e., Δ_2 in (3) or $\Delta_2,\ldots,\Delta_k$ in Definition 2.1). This requires leveraging domain knowledge and additional information to answer the question *what part of the measured outcomes is not attributed to the treatment intervention?* In this section we illustrate three applications of ANA where bivariate metric decompositions arise by engineering an ANC (Section 3.1), identifying an ANC that arises naturally in a funnel decomposition (Section 3.2), and defining an ANC in the context of a surrogate metric (Section 3.3). In each of these sections we find that ANA improves sensitivity and increases the number of stat. sig. results. Through A/A tests and type I error control, in Section 3.4 we demonstrate that ANA does not *arbitrarily* increase sensitivity, it increases sensitivity to non-null effects only.

3.1 Engineering an Approximately Null Component via Counterfactual Reasoning

We applied approximately null augmentation to 39 early-stage ranking experiments at Airbnb. The goal of these experiments was to compare two versions of the ranking algorithm that determines the order of displayed search results. These early-stage experiments run for roughly 1 week taking a small percentage of total traffic. The main target metric of interest is *bookings per guest*. For each search, a user is given the ranked results which determines both (i) the list of results shown in the feed view and on the map view, and (ii) the order of the results listed in the feed view. See Figure 1 for an example of the feed view and the map view together in the desktop browser experience. In the iOS and Android apps, users can switch between the two views.

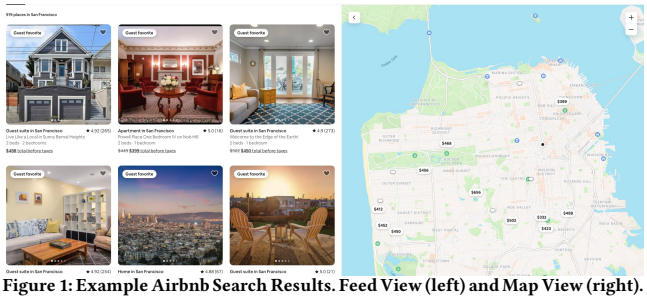

Figure 1: Example Airbnb Search Results. Feed View (left) and Map View (right).

To construct the ANC, we leverage counterfactual ranking results. That is, for treated users (those for whom the treatment ranker generated their ranked feeds and corresponding map view), we also compute the ranked list that *would have* been shown to them if they were assigned to the control group. For control users we similarly computed the ranked list that would have been shown to them had they been assigned to the treatment group. We then construct the approximately null component as described in the following steps:

(1) For each booking conversion, we used attribution logic to attribute the booking to click actions from various search result pages. The attribution is additive such that the sum of the attributed values is 1 for every booking. In this way, the attributed values provide information on the relative importance of various click actions on the booking. Attribution methods are an important research area in their own right. See Deng et al. [7] for more discussion of the attribution logic and the method used in our application.
(2) We then select a subset of attributed search result clicks leading to every booking. A click is selected if:
    (a) the clicked result is ranked among the top 4 results by both the treatment and the control ranker, with a ranked position difference no more than 2, or

(b) both treatment and control rankers show the clicked result on the map view, and the click happens only on the map (i.e., there was no click on the feed view).

(3) For each booking, define the ANC (Component 2, Δ_2) as the sum of attributed values for all selected clicks from the last step. The signal component (Component 1, Δ_1) is straightforwardly defined as the complement of the decomposition such that the two components sum to 1. In other words, Component 1 is the sum of all attributed values from clicks that were *not* included in the last step.

(4) Aggregate decomposed bookings to the user level and then to the treatment/control group level.

The heuristics behind this process can be explained as follows. The criterion in 2(a) considers cases where the booked listing was highly ranked by both the factual and the counterfactual rankers. This kind of booked listing is considered "easy" in the sense that any sensible ranker would put this listing within the first few results. Moreover, we require the ranked position difference to be no more than 2 to further restrict the proximity of the two rankers on this booked listing's position. The intuition is that this type of booking would have happened regardless of which ranker was used, and thus the treatment effect should be approximately null. The criterion in 2(b) is based on the assumption that if search results are clicked on the map (i.e., not the feed view), and both rankers put this listing on the map, then the booking would happen regardless of which ranker was used. Thus the treatment effect for such clicks should be approximately null. Note that these heuristics ignore second order effects like the possibility that a user's booking behavior also depends on the the whole set of the results, not just the ranked position of the booked listing. However, we do not aim nor do we need to guarantee zero treatment effect on Component 2, we only aspire to limit the treatment effect on this component so it has a much smaller effect compared to Component 1.

The effect covariance $\mathbf{\Lambda}$ and the average covariance of the noise $\boldsymbol{\varepsilon}$ were estimated to be (after scaling by the same constant)

$$\mathbf{\Lambda} = \begin{pmatrix} 3.479 & -0.979 \\ -0.979 & 0.672 \end{pmatrix} \quad \text{and} \quad \overline{\Sigma} = \begin{pmatrix} 0.779 & 0.162 \\ 0.162 & 4.096 \end{pmatrix}.$$

We find the approximately null component Δ_2 displays a noise variance 5 times larger than the signal component Δ_1 (4.096 vs. 0.779), while the variance of the treatment effects for Δ_2 is less than 1/5 of that of Δ_1 (0.672 vs. 3.479). This means the SNR of Component 2 is much smaller than that of Component 1 (see equation (5)).

Table 1 summarizes these results and shows that the SNRs of the two components differ by almost a factor of 30. Component 1's SNR is also more than 10 times greater than the SNR of the original metric without decomposition. This suggests that if Component 2's effect is truly much smaller than Component 1, Δ_1 alone can be an estimate for the target metric's treatment effect, with smaller noise variance and hence much greater statistical power. Indeed, among the 39 early-stage experiments, Component 2 was stat. sig. at a 5% level only twice (i.e., 5.1% of the time). This is very close to the 5% significance level, indicating that Component 2 is approximately null. Component 1, on the other hand, was stat. sig. in 13 of the 39 experiments. Without metric decomposition, the booking metric was only stat. sig. 6 times.

We applied the five versions of ANA adjustment discussed in Section 2.2 in this example: ANA to maximize correlation (denoted ANA_c), to minimize mean squared error (denoted ANA_e), to minimize variance (denoted ANA_v), to maximize the expected squared Z-score (denoted ANA_z) and to maximize power (denoted ANA_p). Table 1 demonstrates that all these adjustment methods yield similar results to analyses using Component 1 alone, though ANA_e has one less stat. sig. result out of 39 experiments. This is because the adjustment coefficient θ for all objectives tends to be relatively small for these experiments. Figure 2 plots the optimal θ values. We see these values range between -0.2 to 0.2. Figure 3 shows the five ANA estimates Δ^*. Despite different objectives, estimates in this application don't differ materially, aligning with the similar test results in Table 1. Finally, Figure 4 compares the variances of these ANA estimators. We see that minimum variance objective (ANA_v) provides the lower bound of what variance is achievable through augmentation. In general, minimizing variance directly could lead to more bias relative to using the high SNR component (Δ_1) alone, since the second component (Δ_2) is only *approximately* null. However, in this application Component 2's SNR is so low (relative to Component 1), it suggests augmentation by Component 2 is essentially a null augmentation.

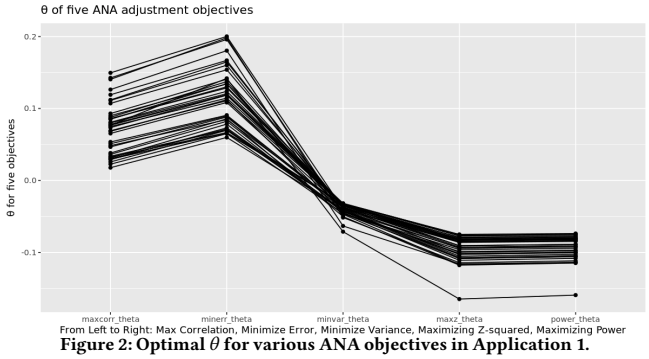

Figure 2: Optimal θ for various ANA objectives in Application 1.

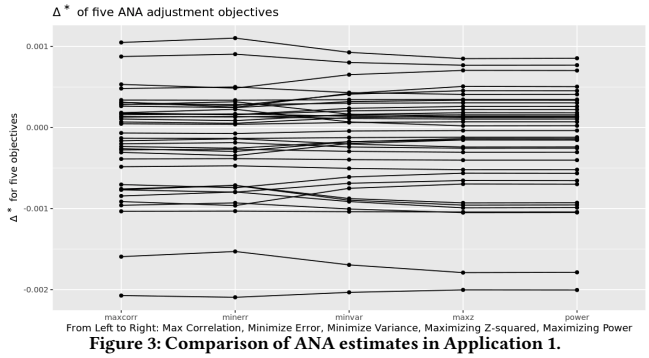

Figure 3: Comparison of ANA estimates in Application 1.

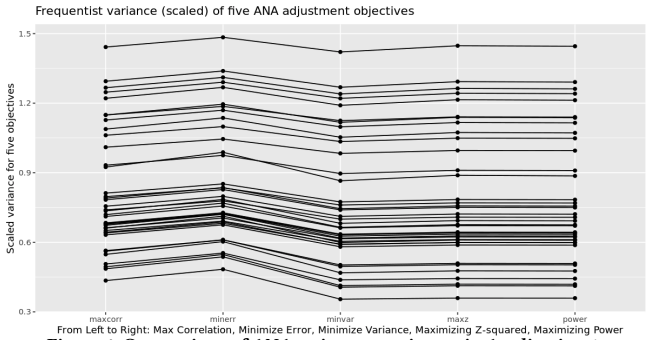

Figure 4: Comparison of ANA estimator variances in Application 1.

	Comp. 1 (Δ_1)	Comp. 2 (Δ_2)	No Decomposition (Δ)	ANA$_c$ (Δ_c^*)	ANA$_e$ (Δ_e^*)	ANA$_v$ (Δ_v^*)	ANA$_z$ (Δ_z^*)	ANA$_p$ (Δ_p^*)
Signal: Var$[\delta]$	3.479	0.672	2.193					
Noise: Var$[\varepsilon]$	0.779	4.096	5.198					
Signal-Noise-Ratio	**4.466**	**0.164**	**0.422**					
Proportion of Stat. Sig.	13/39 (33.3%)	2/39 (5.1%)	6/39 (15.4%)	13/39 (33.3%)	12/39 (30.8%)	13/39 (33.3%)	13/39 (33.3%)	13/39 (33.3%)

Table 1: Results of Application 1 (Engineering an ANC via Counterfactual Reasoning).

3.2 A Metric with Natural Multiplicative Decomposition

In the last section we exploited domain knowledge and context-specific information to engineer an approximately null component. Here we consider a context in which an ANC arises naturally in a conversion funnel where the treatment intervention mainly impacts one step of the funnel and has close to zero impact on the other steps. To illustrate this, we study the metric *nights per guest* which quantifies the number of nights booked per guest. This metric naturally decomposes into *nights per booking*, and *bookings per guest*. As explained in Section 1.1, a multiplicative decomposition of percent treatment effects can be treated as an additive decomposition when the lifts are expected to be small. In this study, we use the decomposition

$$\Delta\%(\text{Nights/Guest}) \approx \Delta\%(\text{Nights/Booking}) + \Delta\%(\text{Bookings/Guest})$$

where $\Delta\%$(Nights/Booking) is the approximately null component (Δ_2), and $\Delta\%$(Bookings/Guest) is the signal component (Δ_1).

We analyze 116 past A/B tests separately with each of the three metrics: nights per guest, nights per booking, and bookings per guest. The results are summarized in Table 2. Generally speaking, the results are very similar to those from the previous application in Table 1. First, Component 2 (nights per booking) has close to 5% empirical stat. sig. rate (5 out of 116) with a very low SNR of 0.014. Second, Component 1 (bookings per guest) has a much larger SNR, and higher empirical stat. sig. rate. (30 out of 116). Analysis with Component 1 also shows better performance than an analysis without decomposition (i.e., analyzing with respect to nights per booking), which has just 11 out of 116 stat. sig. results. Further, analyses with all five ANA adjustments give similar results to analyses with Component 1 alone. As with the previous application, this is because the optimal θ values (though not pictured here) are close to 0.

3.3 Adjustment of a Surrogate Metric

As discussed in Section 2.3, an important area of related work that also seeks to increase metric sensitivity is to build a surrogate or proxy metric. The goal is to use one or more candidate metrics to form an index to better track the treatment effect of a metric of interest, or construct a model-based prediction for the metric of interest using a set of predictors [2, 7, 15, 28].

Let S be a surrogate metric for a metric M, then this implies a decomposition

$$\Delta(M) = \Delta(S) + \Delta(R)\,,$$

where $R = M - S$ is the residual. Therefore any surrogate metric is always associated with a decomposition, and the surrogate metric framework can therefore be seen as a special case of metric decomposition. Moreover, if a surrogate metric is unbiased, then

$$\mathrm{E}[\Delta(M)] = \mathrm{E}[\Delta(S)]\,,$$

and $\mathrm{E}[\Delta(R)] = 0$. This means a surrogate without bias is also a decomposition with null augmentation![4] However, in practice we don't expect to achieve an unbiased surrogate and instead aim for small bias; this of course lends itself well to the benefits of approximately null augmentation. Thus, we advocate for the general use of the metric decomposition and ANA framework for two reasons:

(1) Defining surrogate metrics and then verifying their small bias is often harder than directly constructing a decomposition that has an approximately null effect. The latter is more straightforward because we can leverage natural decompositions from conversion funnels or leverage domain knowledge and counterfactual information, as demonstrated in the previous two applications.
(2) Even when a surrogate metric is available, we can always apply an ANA adjustment to the decomposition implied by the surrogate and its residual.

This latter point is illustrated in this application where we took a surrogate metric for booking-per-guest and studied its decomposition across 133 experiments. This surrogate metric utilizes a set of upper funnel signals to predict a future conversion. It is known that this type of surrogate metric is only unbiased under strong surrogacy assumptions [2]. Table 3 summarizes the results. Comparing to the previous two applications, one distinct difference is that Component 2 (the residual) has a noticeably greater SNR (1.044), and a higher (12.8%, 17 out of 133) stat. sig. rate than the nominal 5% significance level. This indicates that Component 2 may still contain some treatment effect not fully captured by the surrogate metric (Component 1). Nevertheless, relative to the original (un-decomposed) metric, the surrogate component has a greater SNR (6.456 vs. 1.794) and higher statistical power (55 out of 133 stat. sig. results compared to 34 out of 133).

ANA in this case would create an adjusted estimate of the form $\Delta(S) + \theta[\Delta(M) - \Delta(S)]$, that pulls the surrogate metric closer to the original metric by the factor of θ. Figure 5 displays the optimal ANA θ across these experiments for each of the 5 objectives discussed in Section 2.2. We see that maximizing correlation and minimizing error resulted in larger θ ranging from 0.4 to 0.8, and a smaller number of stat. sig. results (45 out of 133). On the other hand, minimizing variance, maximizing expected squared Z-score, and maximizing power resulted in smaller values of θ (ranging between 0 and 0.3) so the ANA estimator is closer to using the surrogate metric (Component 1, Δ_1) directly. Interestingly, these latter 3 objectives resulted in slightly more stat. sig. results than the surrogate metric alone.

3.4 ANA in A/A Tests

In the previous three subsections, we have celebrated an increase in the number of stat. sig. results when using ANA versus an undecomposed analysis. However, it's important to consider whether this increase in stat. sig. results coincides with an increase in type I error.

[4] We can also use multiplicative decomposition with $R = M/S$ and $\Delta\%(M) = \Delta\%(S) + \Delta\%(R)$.

	Comp. 1 (Δ_1)	Comp. 2 (Δ_2)	No Decomposition (Δ)	ANA$_c$ (Δ_c^*)	ANA$_e$ (Δ_e^*)	ANA$_v$ (Δ_v^*)	ANA$_z$ (Δ_z^*)	ANA$_p$ (Δ_p^*)
Signal: Var$[\delta]$	6.508	0.074	5.198					
Noise: Var$[\varepsilon]$	2.810	5.321	8.133					
Signal-Noise-Ratio	**2.316**	**0.014**	**0.639**					
Proportion of Stat. Sig.	30/116 (25.9%)	5/116 (4.3%)	11/116 (9.5%)	30/116 (25.9%)	30/116 (25.9%)	30/116 (25.9%)	30/116 (25.9%)	30/116 (25.9%)

Table 2: Results of Application 2 (Natural Multiplicative Decomposition).

	Comp. 1 (Δ_1)	Comp. 2 (Δ_2)	No Decomposition (Δ)	ANA$_c$ (Δ_c^*)	ANA$_e$ (Δ_e^*)	ANA$_v$ (Δ_v^*)	ANA$_z$ (Δ_z^*)	ANA$_p$ (Δ_p^*)
Signal: Var$[\delta]$	1.011	0.352	0.585					
Noise: Var$[\varepsilon]$	0.157	0.337	0.326					
Signal-Noise-Ratio	**6.456**	**1.044**	**1.794**					
Proportion of Stat. Sig.	55/133 (41.4%)	17/133 (12.8%)	34/133 (25.6%)	45/133 (33.8%)	45/133 (33.8%)	57/133 (42.9%)	58/133 (43.6%)	58/133 (43.6%)

Table 3: Results of Application 3 (Adjustment of a Surrogate Metric).

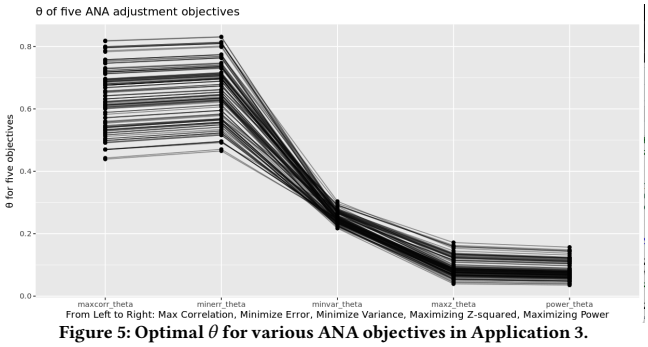

Figure 5: Optimal θ for various ANA objectives in Application 3.

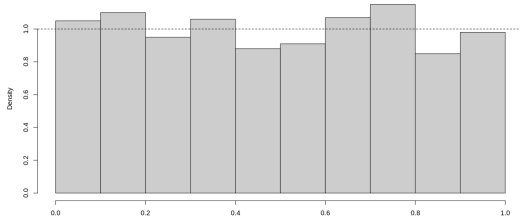

Figure 6: Empirical distribution of p-values of an ANA estimator from 1000 A/A tests.

In this section we emphasize that ANA does not increase sensitivity in general, it increases sensitivity to *non-null effects*. To demonstrate that ANA does not inflate type I error, we simulated 1000 A/A tests (where the treatment effect is truly null) and performed an ANA-based analysis on each. In particular, we randomly split one experiment's data into pseudo treatment and control groups 1000 times and computed p-values for $H_0 : \delta = 0$ when the estimator is taken to be Δ_c^* (i.e., ANA to maximize correlation). Figure 6 shows that the p-values for these tests were uniformly distributed as expected. The empirical proportions of p-values less than 5% and 10% were respectively 0.05 and 0.105 and hence close to nominal. Though not shown here, other augmentations yielded similar behavior. This should provide assurance that ANA is not arbitrarily increasing the number of stat. sig. results; it is instead increasing sensitivity to truly non-null effects.

4 Bayesian View on Metric Decomposition

Here we elaborate on the Bayesian motivation for metric decomposition. In Section 4.1 we demonstrate theoretically and through an example that metric decomposition reduces posterior variance. And in Section 4.2, we use simulation to explore the circumstances under which the variation reduction elicited by decomposition is large or small.

4.1 Posterior Variance Reduction

The methodology in Section 2, which is predicated on the random effects model (4), is closely related to a Bayesian analysis where we posit a prior distribution for δ and perform inference via posterior analyses [6, 10, 14, 16, 24, 29]. It is well-known that the Bayesian posterior mean will shrink the observed frequentist point estimate towards the *global* mean of the prior, where the shrinkage factor is related to the signal-to-noise ratio. Given a metric decomposition with SNR disparity, the Bayesian posterior mean should shrink each component very differently, resulting in a posterior mean that depends more on high SNR components. Of interest is to investigate whether this new posterior distribution exhibits reduced posterior variance.

We prove here that when we assume $\boldsymbol{\delta}$ has a multivariate normal prior with covariance matrix $\boldsymbol{\Lambda}$, at least for the two-component case, the posterior variance of δ conditioned on the bivariate vector $\boldsymbol{\Delta}$ cannot exceed the posterior variance of δ conditioned only on the univariate Δ. This is also true for the general $k > 2$ case when the noise covariance matrices Σ and $\boldsymbol{\Lambda}$ are co-linear. These results are established in Theorem 1 below.

Theorem 1. *Metric decomposition naturally leads to variance reduction under the Bayesian framework with a Gaussian prior for $\boldsymbol{\delta}$. The posterior variances of δ conditioned on $\boldsymbol{\Delta} = (\Delta_1, \Delta_2)$ and Δ are respectively*

$$\mathrm{Var}[\delta|\boldsymbol{\Delta}] = \mathbf{1}_k^\intercal(\boldsymbol{\Lambda} - \boldsymbol{\Lambda}(\boldsymbol{\Lambda}+\Sigma)^{-1}\boldsymbol{\Lambda})\mathbf{1}_k\,,$$

$$\mathrm{Var}[\delta|\Delta] = \frac{\mathbf{1}_k^\intercal\boldsymbol{\Lambda}\mathbf{1}_k \times \mathbf{1}_k^\intercal\Sigma\mathbf{1}_k}{\mathbf{1}_k^\intercal(\boldsymbol{\Lambda}+\Sigma)\mathbf{1}_k}\,.$$

When $k = 2$, the posterior variance of $\delta = \delta_1 + \delta_2$ under bivariate decomposition cannot exceed the univariate posterior variance, i.e.,

$$\mathrm{Var}[\delta|\boldsymbol{\Delta}] \le \mathrm{Var}[\delta|\Delta]. \tag{12}$$

When $k \ge 2 \in \mathbb{N}$, the above inequality holds strictly when $\Sigma = q\boldsymbol{\Lambda}$ for some scalar constant $q \in \mathbb{R}$.

The proof is provided in Appendix B. Even though we have not proved the inequality in (12) for general k (without the strong collinearity assumption), we conjecture the result holds under much milder assumptions and leave this investigation for future theoretical development.

Next we demonstrate this posterior variance reduction in the context of Application 1 from Section 3.1. The results in that section

corresponded to frequentist analyses, but we also analyzed each of the 39 experiments from the Bayesian perspective, in line with Theorem 1. As the theory suggests, the left panel in Figure 7 demonstrates that with a Gaussian prior, the posterior variance is greatly reduced with the bivariate decomposition compared to the univariate analysis without decomposition. Furthermore, the right panel in Figure 7 also illustrates that the Bayesian Z-score (posterior mean divided by posterior standard deviation) tends to have a larger magnitude under the bivariate decomposition. However, this result does not hold uniformly; 1 of the 39 experiments has a larger Z-score with the non-decomposed analysis. Thus, Theorem 1 guarantees a variance reduction, but it does not guarantee an increase in power; sometimes the reduction in the size of the posterior mean may be substantial enough to offset the variance reduction achieved with the decomposition.

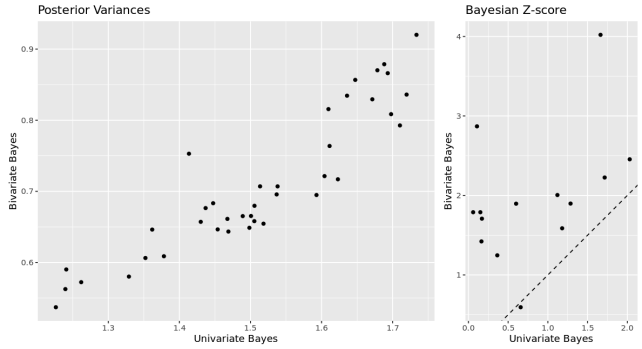

Figure 7: Comparison of bivariate decomposed vs. univariate non-decomposed models with respect to posterior variances (left) and Bayesian Z-scores (right).

4.2 Simulation Illustrating the Benefit of Decomposition

In Section 4.1 we established that the posterior variance of δ when conditioned on $\mathbf{\Delta}$ cannot be larger than when conditioned on Δ. However, we did not explore what variation reduction is achievable by decomposition, nor did we explore when the variation reduction is negligible. The numerical study presented in this section provides insights into this. Here we use a more helpful parameterization of $\mathbf{\Lambda}$ and $\mathbf{\Sigma}$:

$$\mathbf{\Lambda} = \lambda_{11}\begin{bmatrix} 1 & \sqrt{K}\rho_\lambda \\ \sqrt{K}\rho_\lambda & K \end{bmatrix} \text{and} \mathbf{\Sigma} = \lambda_{11}\begin{bmatrix} 1/S_1 & \sqrt{K/(S_1S_2)}\rho_\Sigma \\ \sqrt{K/(S_1S_2)}\rho_\Sigma & K/S_2 \end{bmatrix}, \tag{13}$$

where $\lambda_{11} = \text{Var}[\delta_1]$, $K = \text{Var}[\delta_2]/\lambda_{11}$, $\rho_\lambda = \text{Corr}[\delta_1, \delta_2]$, $\rho_\Sigma = \text{Corr}[\varepsilon_1, \varepsilon_2]$, and $S_1 = \text{Var}[\delta_1]/\text{Var}[\varepsilon_1]$ and $S_2 = \text{Var}[\delta_2]/\text{Var}[\varepsilon_2]$ are signal-to-noise ratios. We can freely vary these parameters and still satisfy the Cauchy-Schwarz inequality.

Here we compare the posterior variances in the decomposed and un-decomposed models for each combination of the following parameter values:

- $K, S_1, S_2 = \{0.01, 0.11, \ldots, 0.91, 1.01, 2, 3, 4, 5\}$
- $\rho_\lambda, \rho_\Sigma = \{-0.975, -0.925, \ldots, 0.975\}$

Because changing the value for λ_{11} just scales $\mathbf{\Lambda}$ and $\mathbf{\Sigma}$ by the same constant, we do not consider it in our simulations. For each of these 5.4×10^6 combinations, we computed the ratio of variances in the posteriors that do and do not account for the bivariate decomposition. As expected, we found the variances to be equal when $\rho_\lambda = \rho_\Sigma$ and $S_1 = S_2$. Under these conditions, $\mathbf{\Sigma} = q\mathbf{\Lambda}$.

Across the 5.4×10^6 parameter combinations, we found that variance reduction is greatest when three conditions are satisfied: the signal-to-noise ratios S_1 and S_2 differ substantially, $|\rho_\lambda|$ is large, and $|\rho_\Sigma|$ is large. To summarize, we visualize the magnitude of the variance reduction factor under various scenarios where these conditions are and are not satisfied. Figure 8 plots the density curves of the variance reduction factor under several scenarios with small and large $|\rho_\lambda|$ and $|\rho_\Sigma|$ when $|S_1 - S_2| > 2$. In all such scenarios, the SNRs differ substantially. As expected, the median reduction factor is largest in the top left plot, where $|\rho_\lambda|$ and $|\rho_\Sigma|$ are large. The plots on the off-diagonals consider scenarios where only one of $|\rho_\lambda|$ or $|\rho_\Sigma|$ is large. Figure 8 suggests that strong correlation between δ_1 and δ_2 is more beneficial than strong correlation between ε_1 and ε_2. The median variance reduction factor is smallest in the bottom right plot, where $|\rho_\lambda|$ and $|\rho_\Sigma|$ are small. Moreover, we find that when SNRs are relatively similar (i.e., $|S_1 - S_2| < 0.1$), analogous plots (See Appenix C) indicate minimal variance reduction, no matter the values of $|\rho_\lambda|$ or $|\rho_\Sigma|$. These results suggest that discrepancies (or the lack thereof) between the SNRs play a greater role than $|\rho_\lambda|$ or $|\rho_\Sigma|$ in variance reduction.

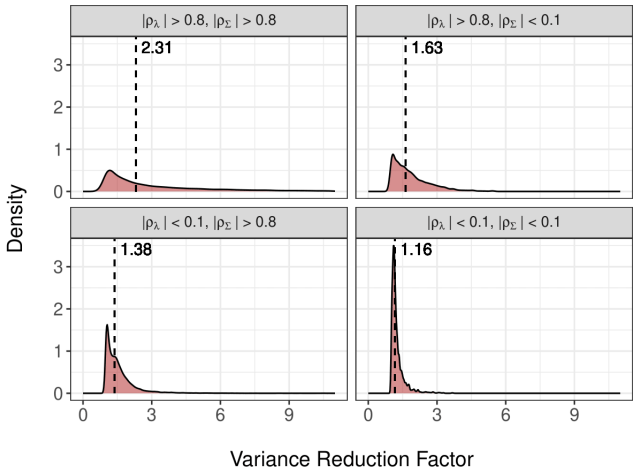

Figure 8: Density curves of the variance reduction factor for several conditions when signal-to-noise ratios differ substantially. Median reduction factors are given by the dashed vertical lines and annotated text.

5 Conclusion & Discussion

In this paper we have proposed metric decomposition as a novel means to improve metric sensitivity in online A/B tests. The idea is premised upon the decomposition of a target metric into two or more components that differ with respect to their signal-to-noise ratios. We show through theory, simulation, and empirical examples that if such a decomposition exists (or can be engineered), sensitivity may be increased via approximately null augmentation (in a frequentist setting) and posterior variance is reduced (in a Bayesian setting). We provide practical guidance for, and discuss the implications of, metric decomposition in both settings. We also contrast it with industry-favorite alternatives like CUPED, and in doing so highlight its broad utility. An important extension to this work would be to next consider sample size determination in both the frequentist or Bayesian contexts; while a boost in sensitivity typically means less data is required for a given analysis, a methodology that determines the smallest sample size required to control various operating characteristics in this context would be of practical value.

References

[1] Soren Asmussen and Peter Glynn. 2008. *Stochastic Simulation.* Springer-Verlag.

[2] Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. 2019. *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely.* Technical Report. National Bureau of Economic Research.

[3] Iavor Bojinov and Somit Gupta. 2022. Online Experimentation: Benefits, Operational and Methodological Challenges, and Scaling Guide. *Harvard Data Science Review* 4, 3 (jul 28 2022). https://hdsr.mitpress.mit.edu/pub/aj31wj81.

[4] Laura Cosgrove, Jen Townsend, and Jonathan Litz. [n. d.]. Deep Dive Into Variance Reduction. https://www.microsoft.com/en-us/research/group/experimentation-platform-exp/articles/deep-dive-into-variance-reduction/.

[5] Tom Cunningham. 2023. Experiment Interpretation and Extrapolation. https://tecunningham.github.io/posts/2023-04-18-experiment-interpretation-extrapolation.html

[6] Alex Deng. 2015. Objective Bayesian Two Sample Hypothesis Testing for Online Controlled Experiments. In *Proceedings of the 24th International Conference on World Wide Web Companion*. 923–928.

[7] Alex Deng, Michelle Du, Anna Matlin, and Qing Zhang. 2023. Variance Reduction Using In-Experiment Data: Efficient and Targeted Online Measurement for Sparse and Delayed Outcomes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3937–3946.

[8] Alex Deng and Victor Hu. 2015. Diluted treatment effect estimation for trigger analysis in online controlled experiments. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. 349–358.

[9] Alex Deng, Ulf Knoblich, and Jiannan Lu. 2018. Applying the Delta Method in Metric Analytics: A Practical Guide with Novel Ideas. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (London, United Kingdom) *(KDD '18)*. ACM, New York, NY, USA, 233–242.

[10] Alex Deng, Yicheng Li, Jiannan Lu, and Vivek Ramamurthy. 2021. On Post-selection Inference in A/B Testing. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2743–2752.

[11] Alex Deng and Xiaolin Shi. 2016. Data-driven metric development for online controlled experiments: Seven lessons learned. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[12] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the 6th ACM WSDM Conference*. 123–132.

[13] Alex Deng, Lo-Hua Yuan, Naoya Kanai, and Alexandre Salama-Manteau. 2023. Zero to hero: Exploiting null effects to achieve variance reduction in experiments with one-sided triggering. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 823–831.

[14] Drew Dimmery, Eytan Bakshy, and Jasjeet Sekhon. 2019. Shrinkage Estimators in Online Experiments. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2914–2922.

[15] Weitao Duan, Shan Ba, and Chunzhe Zhang. 2021. Online Experimentation with Surrogate Metrics: Guidelines and a Case Study. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 193–201.

[16] Bradley Efron. 2010. *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction.* Cambridge University Press.

[17] Michael R Elliott, Anna SC Conlon, Yun Li, Nico Kaciroti, and Jeremy MG Taylor. 2015. Surrogacy marker paradox measures in meta-analytic settings. *Biostatistics* 16, 2 (2015), 400–412.

[18] Andrew Gelman and John Carlin. 2014. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 9, 6 (2014), 641–651.

[19] Andrew Gelman and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press.

[20] Somit Gupta et al. 2019. Top Challenges from the First Practical Online Controlled Experiments Summit. *SIGKDD Explor. Newsl.* 21, 1 (May 2019), 20–35.

[21] Ron Kohavi, Alex Deng, and Lukas Vermeer. 2022. A/B Testing Intuition Busters: Common Misunderstandings in Online Controlled Experiments. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3168–3177.

[22] Ron Kohavi, Diane Tang, and Ya Xu. 2020. *Trustworthy online controlled experiments: A practical guide to a/b testing.* Cambridge University Press.

[23] Nicholas Larsen, Jonathan Stallrich, Srijan Sengupta, Alex Deng, Ron Kohavi, and Nathaniel T Stevens. 2023. Statistical challenges in online controlled experiments: A review of a/b testing methodology. *The American Statistician* (2023), 1–15.

[24] Sareh Nabi, Houssam Nassif, Joseph Hong, Hamed Mamani, and Guido Imbens. 2022. Bayesian meta-prior learning using Empirical Bayes. *Management Science* 68, 3 (2022), 1737–1755.

[25] Art B Owen. 2013. Monte Carlo theory, methods and examples. (2013).

[26] Alexander Peysakhovich and Dean Eckles. 2018. Learning causal effects from many randomized experiments using regularized instrumental variables. In *Proceedings of the 2018 World Wide Web Conference*. 699–707.

[27] Daniel Ting and Kenneth Hung. 2023. On the Limits of Regression Adjustment. *arXiv preprint arXiv:2311.17858* (2023).

[28] Nilesh Tripuraneni, Lee Richardson, Alexander D'Amour, Jacopo Soriano, and Steve Yadlowsky. 2023. Choosing a Proxy Metric from Past Experiments. *arXiv preprint arXiv:2309.07893* (2023).

[29] Runzhe Wan, Yu Liu, James McQueen, Doug Hains, and Rui Song. 2023. Experimentation platforms meet reinforcement learning: Bayesian sequential decision-making for continuous monitoring. *arXiv preprint arXiv:2304.00420* (2023).

Appendix

A ANA Derivations

Consider ANA estimators $\Delta^* := \boldsymbol{c}^\top \boldsymbol{\Delta}$, where $\boldsymbol{c}^\top \boldsymbol{e}_1 = 1$ for the standard basis vector $\boldsymbol{e}_1$. Let $\boldsymbol{c}_*$ be the final $k-1$ components of $\boldsymbol{c}$ and $\boldsymbol{\Lambda}_*$ be the $(k-1)\times(k-1)$ submatrix of $\boldsymbol{\Lambda}$ corresponding to $(\delta_2,\ldots,\delta_k)$. Let Σ_* and be the $(k-1)\times(k-1)$ submatrix of Σ corresponding to $(\varepsilon_2,\ldots,\varepsilon_k)$. Let $\Sigma_1 = (\Sigma_{12},\ldots,\Sigma_{1k})$.

Minimizing Mean Squared Error. We have that

$$\begin{aligned}
&\mathrm{E}\left[(\delta-\Delta^*)^2\right] \\
&= \mathrm{E}\left[\left(\mathbf{1}_k^\top \boldsymbol{\delta} - \boldsymbol{c}^\top(\boldsymbol{\delta}+\boldsymbol{\varepsilon})\right)^2\right] \\
&= \mathrm{E}\left[\left((\mathbf{1}_k-\boldsymbol{c})^\top \boldsymbol{\delta} - \boldsymbol{c}^\top\boldsymbol{\varepsilon}\right)^2\right] \\
&= \mathrm{E}\left[\left((\mathbf{1}_k-\boldsymbol{c})^\top \boldsymbol{\delta}\boldsymbol{\delta}^\top(\mathbf{1}_k-\boldsymbol{c}) - 2(\mathbf{1}_k-\boldsymbol{c})^\top\boldsymbol{\delta}\boldsymbol{\varepsilon}^\top\boldsymbol{c} + \boldsymbol{c}^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{c}\right] \\
&= (\mathbf{1}_k-\boldsymbol{c})^\top\boldsymbol{\Lambda}(\mathbf{1}_k-\boldsymbol{c}) + \boldsymbol{c}^\top\Sigma\boldsymbol{c} \\
&= (\mathbf{1}_{k-1}-\boldsymbol{c}_*)^\top\boldsymbol{\Lambda}_*(\mathbf{1}_{k-1}-\boldsymbol{c}_*) + \Sigma_{11} + 2\boldsymbol{c}_*^\top\Sigma_1 + \boldsymbol{c}_*^\top\Sigma_*\boldsymbol{c}_*.
\end{aligned} \tag{A.1}$$

The penultimate step follows because $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ are independent, and the final equality holds because $\boldsymbol{c}^\top\boldsymbol{e}_1 = 1$. The derivative of (A.1) with respect to $\boldsymbol{c}_*$ is

$$\frac{\partial}{\partial \boldsymbol{c}_*}\mathrm{E}\left[(\delta-\Delta^*)^2\right] = -2\boldsymbol{\Lambda}_*(\mathbf{1}_{k-1}-\boldsymbol{c}_*) + 2\Sigma_1 + 2\Sigma_*\boldsymbol{c}_*. \tag{A.2}$$

The value for $\boldsymbol{c}_*$ that equates the expression in (A.2) to $\mathbf{0}$ and hence minimizes mean squared error is

$$\boldsymbol{c}_* = \left[\boldsymbol{\Lambda}_* + \Sigma_*\right]^{-1}\left[\boldsymbol{\Lambda}_*\mathbf{1}_{k-1} - \Sigma_1\right]. \tag{A.3}$$

The value of θ given in (6) is obtained as a special case when $k=2$ and $\boldsymbol{c} = (1,\theta)^\top$. □

Maximizing Correlation. We have that

$$\begin{aligned}
\mathrm{Corr}\left(\delta,\Delta^*\right) &= \frac{\mathrm{Cov}(\mathbf{1}_k^\top\boldsymbol{\delta}, \boldsymbol{c}^\top\boldsymbol{\Delta})}{\sqrt{\mathrm{Var}(\mathbf{1}_k^\top\boldsymbol{\delta})}\sqrt{\mathrm{Var}(\boldsymbol{c}^\top\boldsymbol{\Delta})}} \\
&= \frac{\mathbf{1}_k^\top\boldsymbol{\Lambda}\boldsymbol{c}}{\sqrt{\mathbf{1}_k^\top\boldsymbol{\Lambda}\mathbf{1}_k}\sqrt{\boldsymbol{c}^\top(\boldsymbol{\Lambda}+\Sigma)\boldsymbol{c}}}.
\end{aligned} \tag{A.4}$$

To maximize (A.4), we take the partial derivative

$$\frac{\partial}{\partial \boldsymbol{c}}\frac{(\mathbf{1}_k^\top\boldsymbol{\Lambda}\boldsymbol{c})^2}{\boldsymbol{c}^\top(\boldsymbol{\Lambda}+\Sigma)\boldsymbol{c}} = \frac{2(\mathbf{1}_k^\top\boldsymbol{\Lambda}\boldsymbol{c})\boldsymbol{\Lambda}\mathbf{1}_k\boldsymbol{c}^\top(\boldsymbol{\Lambda}+\Sigma)\boldsymbol{c} - 2(\boldsymbol{\Lambda}+\Sigma)\boldsymbol{c}(\mathbf{1}_k^\top\boldsymbol{\Lambda}\boldsymbol{c})^2}{(\boldsymbol{c}^\top(\boldsymbol{\Lambda}+\Sigma)\boldsymbol{c})^2}. \tag{A.5}$$

Equating the numerator of (A.5) to $\mathbf{0}$ prompts the following result:

$$(\boldsymbol{\Lambda}+\Sigma)^{-1}\boldsymbol{\Lambda}\mathbf{1}_k = \frac{\mathbf{1}_k^\top\boldsymbol{\Lambda}\boldsymbol{c}}{\boldsymbol{c}^\top(\boldsymbol{\Lambda}+\Sigma)\boldsymbol{c}} \times \boldsymbol{c}. \tag{A.6}$$

Since the scaling constant to the left of the $\times$ sign in (A.6) is applied to each component of $\boldsymbol{c}$, we have that $S^\top\mathbf{1}_k \propto \boldsymbol{c}$. To enforce the constraint that $\boldsymbol{c}^\top\boldsymbol{e}_1 = 1$, we require

$$\boldsymbol{c} = \frac{1}{\boldsymbol{e}_1^\top S^\top \mathbf{1}_k} S^\top \mathbf{1}_k. \tag{A.7}$$

This is the augmentation that maximizes correlation. The value of θ given in (7) is obtained as a special case when $k=2$ and $\boldsymbol{c} = (1,\theta)^\top$. □

Minimizing Error Variance. We have that

$$\begin{aligned}
\mathrm{Var}\left[\boldsymbol{c}^\top\boldsymbol{\varepsilon}\right] &= \boldsymbol{c}^\top\Sigma\boldsymbol{c} \\
&= \Sigma_{11} + 2\boldsymbol{c}_*^\top\Sigma_1 + \boldsymbol{c}_*^\top\Sigma_*\boldsymbol{c}_*
\end{aligned} \tag{A.8}$$

The final equality holds because $\boldsymbol{c}^\top \boldsymbol{e}_1 = 1$. The derivative of (A.8) with respect to $\boldsymbol{c}_*$ is

$$\frac{\partial}{\partial \boldsymbol{c}_*} \text{Var}[\boldsymbol{c}^\top \boldsymbol{\varepsilon}] = 2(\Sigma_1 + \Sigma_* \boldsymbol{c}_*). \tag{A.9}$$

The value for $\boldsymbol{c}_*$ that equates the expression in (A.9) to $\mathbf{0}$ and hence minimizes the error variance is

$$\boldsymbol{c}_* = -\Sigma_*^{-1}\Sigma_1. \tag{A.10}$$

The value of θ given in (8) is obtained as a special case when $k=2$ and $\boldsymbol{c} = (1,\theta)^\top$. □

Maximizing Expected Squared Z-Score. The expected squared Z-score has the form:

$$\begin{aligned} \frac{\text{Var}[\Delta^*]}{\text{Var}[\boldsymbol{c}^\top \boldsymbol{\varepsilon}]} &= \frac{\boldsymbol{c}^\top (\boldsymbol{\Lambda}+\Sigma)\boldsymbol{c}}{\boldsymbol{c}^\top \Sigma \boldsymbol{c}} \\ &= \frac{[\Sigma^{1/2}\boldsymbol{c}]^\top [\Sigma^{-1/2}(\boldsymbol{\Lambda}+\Sigma)\Sigma^{-1/2}][\Sigma^{1/2}\boldsymbol{c}]}{[\Sigma^{1/2}\boldsymbol{c}]^\top [\Sigma^{1/2}\boldsymbol{c}]}. \end{aligned} \tag{A.11}$$

Let $\boldsymbol{x} = \Sigma^{1/2}\boldsymbol{c}$, and (A.11) is a Rayleigh quotient maximized when $\boldsymbol{x}$ is any multiple of the first eigenvector of the matrix $\Sigma^{-1/2}(\boldsymbol{\Lambda}+\Sigma)\Sigma^{-1/2}$. Let $\boldsymbol{x}^*$ be this eigenvector. Then it is easy to see

$$\boldsymbol{c} = \frac{1}{\boldsymbol{e}_1^\top \Sigma^{-1/2} \boldsymbol{x}^*} \Sigma^{-1/2} \boldsymbol{x}^*. \tag{A.12}$$

The value of θ given in (9) is obtained as a special case when $k=2$ and $\boldsymbol{c} = (1,\theta)^\top$. □

Maximizing Power. For a specific value $\dot{\boldsymbol{\delta}} = (\dot{\delta}_1,\ldots,\dot{\delta}_k^\top)$ from the $\boldsymbol{\delta}$ distribution specified by model (4), the test statistic for testing $H_0 : \delta = 0$ is given by

$$\frac{\boldsymbol{c}^\top \dot{\boldsymbol{\delta}}}{\sqrt{\boldsymbol{c}^\top \Sigma \boldsymbol{c}}} = \frac{\dot{\delta}_1 + \boldsymbol{c}_*^\top \dot{\boldsymbol{\delta}}_*}{\sqrt{\Sigma_{11} + 2\boldsymbol{c}_*^\top \Sigma_1 + \boldsymbol{c}_*^\top \Sigma_* \boldsymbol{c}_*}} \tag{A.13}$$

where $\dot{\boldsymbol{\delta}}_*$ is the final $k-1$ components of $\dot{\boldsymbol{\delta}}$. This equality holds because $\boldsymbol{c}^\top \boldsymbol{e}_1 = 1$. The derivative of (A.13) with respect to $\boldsymbol{c}_*$ is

$$\frac{\dot{\boldsymbol{\delta}}_*^\top (\Sigma_{11} + 2\boldsymbol{c}_*^\top \Sigma_1 + \boldsymbol{c}_*^\top \Sigma_* \boldsymbol{c}_*) - (\dot{\delta}_1 + \boldsymbol{c}_*^\top \dot{\boldsymbol{\delta}}_*)(\Sigma_1^\top + \boldsymbol{c}_*^\top \Sigma_*)}{(\Sigma_{11} + 2\boldsymbol{c}_*^\top \Sigma_1 + \boldsymbol{c}_*^\top \Sigma_* \boldsymbol{c}_*)^{3/2}} \tag{A.14}$$

The value for $\boldsymbol{c}_*$ that equates the expression in (A.14) to $\mathbf{0}$ and hence maximizes power is

$$\boldsymbol{c}_* = -\left[\Sigma_1 \dot{\boldsymbol{\delta}}_*^\top - \dot{\delta}_1 \Sigma_*\right]^{-1} \left[\Sigma_{11} \dot{\boldsymbol{\delta}}_* - \dot{\delta}_1 \Sigma_1\right]. \tag{A.15}$$

The value of θ given in (10) is obtained as a special case when $k=2$ and $\boldsymbol{c} = (1,\theta)^\top$. □

B Proof of Theorem 1

For this proof, we use the following parameterization for $\boldsymbol{\Lambda}$ and Σ:

$$\boldsymbol{\Lambda} = \begin{bmatrix} L^2 & L\sqrt{\lambda_{22}}\rho_\lambda \\ L\sqrt{\lambda_{22}}\rho_\lambda & \lambda_{22} \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}.$$

That is, $L = \sqrt{\lambda_{11}}$. This parameterization allows us to freely vary L across $\mathbb{R}^+$ while satisfying the Cauchy-Schwarz inequality. We now show that each side of the inequality in (12) can be expressed as the ratio of two quadratic functions of L.

For the left side of (12), we show that $\mathbf{1}_2^\top (\boldsymbol{\Lambda} - \boldsymbol{\Lambda}(\boldsymbol{\Lambda}+\Sigma)^{-1}\boldsymbol{\Lambda})\mathbf{1}_2$ takes the form

$$\frac{b_1 L^2 + b_2 L + b_3}{b_4 L^2 + b_5 L + b_6}. \tag{B.1}$$

Through simple algebra, we can show that $b_1 = \lambda_{22}(1-\rho_\lambda^2)(\Sigma_{11} + 2\Sigma_{12} + \Sigma_{22}) + \Sigma_{11}\Sigma_{22} - \Sigma_{12}^2$, $b_2 = 2\sqrt{\lambda_{22}}\rho_\lambda(\Sigma_{11}\Sigma_{22} - \Sigma_{12}^2)$, and $b_3 = \lambda_{22}(\Sigma_{11}\Sigma_{22} - \Sigma_{12}^2)$. The denominator of (B.1) is the determinant of $\boldsymbol{\Lambda}+\Sigma$; it takes the form $b_4 L^2 + b_5 L + b_6$, where $b_4 = \lambda_{22}(1-\rho_\lambda^2) + \Sigma_{22}$, $b_5 = -2\sqrt{\lambda_{22}}\rho_\lambda \Sigma_{12}$, and $b_6 = \Sigma_{11}\lambda_{22} + \Sigma_{11}\Sigma_{22} - \Sigma_{12}^2$.

The algebra is simpler to show the right side of (12) takes the form

$$\frac{b_7 L^2 + b_8 L + b_9}{b_{10} L^2 + b_{11} L + b_{12}}. \tag{B.2}$$

That numerator $\mathbf{1}_2^\top \boldsymbol{\Lambda} \mathbf{1}_2 \times \mathbf{1}_2^\top \Sigma \mathbf{1}_2$ is such that $b_7 = \Sigma_{11} + 2\Sigma_{12} + \Sigma_{22}$, $b_8 = 2\sqrt{\lambda_{22}}\rho_\lambda(\Sigma_{11} + 2\Sigma_{12} + \Sigma_{22})$, and $b_9 = \lambda_{22}(\Sigma_{11} + 2\Sigma_{12} + \Sigma_{22})$. That denominator $\mathbf{1}_2^\top (\boldsymbol{\Lambda} + \Sigma)\mathbf{1}_2$ can be expressed as $b_{10}L^2 + b_{11}L + b_{12}$, where $b_{10} = 1$, $b_{11} = 2\sqrt{\lambda_{22}}\rho_\lambda$, and $b_{12} = \lambda_{22} + \Sigma_{11} + 2\Sigma_{12} + \Sigma_{22}$. The denominator of (B.1) must be non-negative due to the Cauchy Schwarz inequality: $b_4 L^2 + b_5 L + b_6 \geq 0$ for all $L \geq 0$. Moreover, the denominator of (B.2) is a variance, so $b_{10}L^2 + b_{11}L + b_{12} \geq 0$ for all $L \geq 0$.

We can therefore cross multiply the fractions in (B.1) and (B.2) to obtain an equivalent inequality to (12) that is a quartic equation of L:

$$aL^4 + bL^3 + cL^2 + dL + e \geq 0, \tag{B.3}$$

where $a = (\Sigma_{12} + \Sigma_{22})^2 \geq 0$, $b = 2\sqrt{\lambda_{22}}\rho_\lambda(\Sigma_{22} - \Sigma_{11})(\Sigma_{12} + \Sigma_{22})$, $c = \lambda_{22}(\rho_\lambda^2(\Sigma_{11} - \Sigma_{22})^2 - 2(\Sigma_{11} + \Sigma_{12})(\Sigma_{12} + \Sigma_{22}))$, $d = 2\lambda_{22}^{3/2}\rho_\lambda(\Sigma_{11} - \Sigma_{22})(\Sigma_{11} + \Sigma_{12})$, and $e = \lambda_{22}^2(\Sigma_{11} + \Sigma_{12})^2$. If (B.3) holds true for all $L \geq 0$, then the bivariate variance of δ cannot exceed the univariate variance (i.e., the inequality in (12) also holds true). It can be shown via algebra that the coefficients in (B.3) satisfy $D = 64a^3e - 16a^2c^2 + 16ab^2c - 16a^2bd - 3b^4 = 0$. This result implies that (B.3) has two double roots. Because the leading coefficient $a \geq 0$, the quartic equation in (B.3) is non-negative for all $L \geq 0$. Theorem 1 follows directly from this result. □

C Additional Simulation Plot

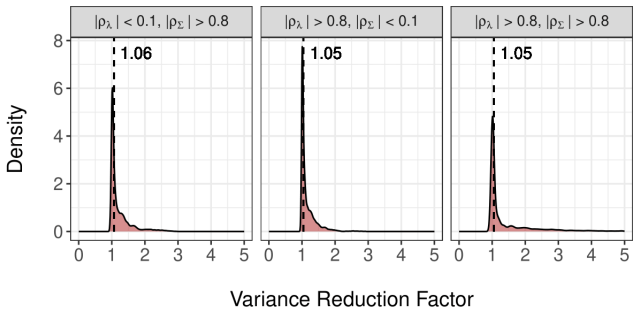

Figure 9: Density curves of the variance reduction factor for several conditions when signal-to-noise ratios do not differ substantially. Median reduction factors are given by the dashed vertical lines and annotated text.