Learning the Covariance of Treatment Effects Across Many Weak Experiments

Aurélien Bibaut
Netflix
Los Gatos, CA, United States

Winston Chou
Netflix
Los Gatos, CA, United States

Simon Ejdemyr
Netflix
Los Angeles, CA, United States

Nathan Kallus
Netflix
New York, NY, United States
Cornell University
New York, NY, United States

ABSTRACT

When primary objectives are insensitive or delayed, experimenters may instead focus on proxy metrics derived from secondary outcomes. For example, technology companies often infer the long-term impacts of product interventions from their effects on short-term user engagement signals. We consider the meta-analysis of many historical experiments to learn the covariance of treatment effects on these outcomes, which can support the construction of such proxies. Even when experiments are plentiful, if treatment effects are weak, the covariance of estimated treatment effects across experiments can be highly biased. We overcome this with techniques inspired by weak instrumental variable analysis. We show that Limited Information Maximum Likelihood (LIML) learns a parameter equivalent to fitting total least squares to a transformation of the scatterplot of treatment effects, and that Jackknife Instrumental Variables Estimation (JIVE) learns another parameter computable from the average of Jackknifed covariance matrices across experiments. We also present a total covariance estimator for the latter estimand under homoskedasticity, which is equivalent to a k-class estimator. We show how these parameters can be used to construct unbiased proxy metrics under various structural models. Lastly, we discuss the real-world application of our methods at Netflix.

CCS CONCEPTS

• **Computing methodologies** → **Learning linear models**; • **Mathematics of computing** → *Multivariate statistics.*

KEYWORDS

surrogate outcomes, meta-analysis, weak instrumental variables

ACM Reference Format:
Aurélien Bibaut, Winston Chou, Simon Ejdemyr, and Nathan Kallus. 2024. Learning the Covariance of Treatment Effects Across Many Weak Experiments. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3637528.3672034

This work is licensed under a Creative Commons Attribution International 4.0 License.

KDD '24, August 25–29, 2024, Barcelona, Spain
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0490-1/24/08.
https://doi.org/10.1145/3637528.3672034

1 INTRODUCTION

Projecting long-term treatment effects from short-term metrics is a ubiquitous problem in experimentation. For example, technology companies seek to optimize insensitive primary metrics (such as habitual usage, subscriber retention, and long-term revenue), but are unable or unwilling to measure treatment effects on these metrics precisely.[1] To address this issue, they optimize a suite of secondary metrics that are associated with the primary metrics, but more sensitive in terms of signal-to-noise and easier to measure in short-term experiments. Under certain assumptions, the treatment effect on a "surrogate index" of multiple secondary metrics yields a precise estimate of the long-term treatment effect [4, 17].

A class of statistical parameters that intuitively relates to the problem of constructing proxy metrics for a primary metric is the covariance matrix of true average treatment effects (ATEs) on primary and secondary metrics in previous experiments, and functions thereof. For example, when constructing weighted indices of secondary outcomes, it is intuitive to consider the Ordinary (OLS) and Total Least Squares (TLS) regression of true ATEs on a primary outcome on true ATEs on the secondary outcomes in the scatterplot of true ATEs over available historical experiments. Here, by true ATE, we mean the unobserved mean on the population, in contrast to the estimated ATE actually observed on the experimental sample.

How these statistical parameters actually connect to the question of proxy metrics or surrogates is the first question we investigate in this paper. In Section 3, we demonstrate that statistical features of the covariance matrix of true ATEs have causal interpretations under different causal models. For instance, under these models, they can support inference on effects of novel treatments on long-term outcomes based on short-term observations.

The second question we study is the estimation of the covariance matrix of the true ATEs when the signal-to-noise ratio is small in each experiment, as is often the case in digital experimentation [8, 16, 20]. The statistical challenges are analogous to the many weak instrumental variables (IVs) setting. One question of interest in the weak instrument literature is whether allowing the number of instruments to diverge while holding their strength fixed yields consistent estimates [15]. Analogously, we show that increasing

[1] For example, they may be interested in the effect of digital platform design on long-term user retention [12], but unwilling to run a sufficiently large experiment for a long time to measure treatment effects on long-term retention.

the number of experiments, even if each maintains a low signal-to-noise ratio, enables consistent estimation of the covariance matrix of true treatment effects. In fact, each of the three methods we study has a weak-instrument estimator counterpart.

Our results demonstrate that we can reliably estimate the covariance matrix of true treatment effects as a parameter at the meta-analytic level, that is, from experiment-level aggregates. In this way, we contribute to an emerging literature on meta-analytic approaches to surrogacy [7, 9, 18], which is particularly relevant given the large number of experiments conducted on modern online experimentation platforms. The sheer volume of historical data from these platforms often makes unit-level analysis computationally challenging, if not prohibitive [19]. Therefore, our methods are not only statistically robust, but also operationally feasible for large-scale experimentation.

The paper is organized as follows:

- In Section 2, we present the data collection process and the statistical parameters.
- In Section 3, we discuss the relationship of the statistical parameters to causal parameters and the construction of proxy metrics.
- In Section 4, we present weak-IV-inspired estimators of the covariance matrix of true treatment effects and the OLS and TLS estimands in the scatterplot of true ATEs.
- In Section 5, we conduct a simulation study to illustrate the performance of our proposed estimators and provide visual intuition on the mechanics of some of our estimators.
- Lastly, in Section 6, we describe the real-world application of our methods to derive a linear surrogate index for experimentation at Netlix.

2 STATISTICAL SETUP AND NOTATION

2.1 Data

We observe N unit-level quadruples $O = (T, A, S, Y)$ where $T \in \{1, \ldots, K\}$ is the experiment index, $A \in \{0, 1\}$ is the treatment arm index, S is a $(G-1)$-dimensional vector of secondary metrics, and Y is the primary metric of interest. We observe $N = Kn$ units divided between K experiments, each of which has two arms with n units.[2] Units in different experiments may be drawn from different superpopulations, but are assigned uniformly completely at random to cells. The observations $O_1, \ldots, O_N \sim O$ are each identically distributed. They may not be independent due to the restriction of having exactly n units in each experiment. Conditional on $T_1, \ldots, T_N$ and $A_1, \ldots, A_N$, any two observations are nonetheless independent.

2.2 Treatment-effect Covariance Matrix

For any $t = 1, \ldots, K$, let $\tau_S(t) = E[S \mid A = 1, T = t] - E[S \mid A = 0, T = t]$ and $\tau_Y(t) = E[Y \mid A = 1, T = t] - E[Y \mid A = 0, T = t]$ be the true ATEs in experiment t on the vector S and on the scalar Y, respectively. Let

$$\Lambda_K = \begin{bmatrix} \Lambda_{YY,K} & \Lambda_{SY,K}^T \\ \Lambda_{SY,K} & \Lambda_{SS,K} \end{bmatrix}$$

with

$$\begin{aligned}
\Lambda_{YY,K} =& \frac{1}{K}\sum_{t=1}^{K} \tau_Y(t)^2 - \left(\frac{1}{K}\sum_{t=1}^{K} \tau_Y(t)\right)^2, \\
\Lambda_{SY,K} =& \frac{1}{K}\sum_{t=1}^{K} \tau_S(t)\tau_Y(t) - \left(\frac{1}{K}\sum_{t=1}^{K} \tau_S(t)\right)\left(\frac{1}{K}\sum_{t=1}^{K} \tau_Y(t)\right), \\
\Lambda_{SS,K} =& \frac{1}{K}\sum_{t=1}^{K} \tau_S(t)\tau_S(t)^\top - \left(\frac{1}{K}\sum_{t=1}^{K} \tau_S(t)\right)\left(\frac{1}{K}\sum_{t=1}^{K} \tau_S(t)\right)^\top.
\end{aligned}$$

That is, Λ_K is the covariance matrix of true ATEs across the primary metric and secondary metrics over the population of experiments. We assume throughout that $\Lambda_{SS,K}$ is positive definite, meaning all its eigenvalues are strictly positive. Note that this requires at least $K \geq G - 1$ tests.

2.3 Statistical Parameters

We consider two statistical parameters. Our first parameter is

$$\theta_1(\Lambda_K) = \Lambda_{SS,K}^{-1}\Lambda_{SY,K}.$$

This is the OLS in the scatterplot of the K true treatment effects. While we discuss the exact causal interpretation of this parameter (which requires structural assumptions) in the next section, intuitively the OLS measures the statistical relationship between the surrogate metrics and the long-term metric [18].

Our second parameter is more complex:

$$\theta_{2,\Psi}(\Lambda_K) = -\gamma_{\Psi,S}(\Lambda_K)/\gamma_{\Psi,Y}(\Lambda_K),$$

where $[\gamma_{\Psi,Y}(\Lambda_K), \gamma_{\Psi,S}(\Lambda_K)^T]^T$ is a generalized eigenvector solving $(\Lambda_K - \kappa\Psi)\gamma = 0$ for the smallest possible $\kappa \geq 0$ for which a solution exists, where Ψ is a given positive definite matrix. The parameter $\theta_{2,\Psi}(\Lambda_K)$ is the TLS on the $\Psi^{-1/2}$-transformed scatterplot of true treatment effects. While this parameter is less intuitive, we show below that it coincides with $\theta_1(\Lambda_K)$ under certain structural assumptions.

We emphasize that these OLS and TLS parameters are defined in terms of only K points in a G-dimensional scatterplot. The location of these points themselves (the true ATEs) are, nonetheless, unknown population quantities. Therefore, $\theta_1(\Lambda_K)$ is a population quantity, meaning it is a function of the unknown population distribution of the data O (that is, of infinite draws of O).

3 CAUSAL MODELS AND RELATIONSHIP TO PROXIES

We now consider various causal models and examine how the statistical parameters $\theta_1(\Lambda_K)$ and $\theta_2(\Lambda_K)$ relate to causal parameters in these models. Equivalently, this section derives the structural assumptions under which these statistical parameters can be given a causal interpretation in terms of the effect of S on Y and thereby used to construct proxy metrics for Y.

[2] While our results extend to the more realistic setting of multiple treatment arms of varying size per experiment, we focus on the two-arm, constant sample size case for ease of exposition.

3.1 Linear Structural Model without Direct Effects

Consider the following linear structural model with experiment-level fixed-effects and unmeasured confounding U between S and Y:

$$\begin{aligned} S &= \mu_S(T) + \pi_S(T)A + \gamma U + \eta \\ Y &= \mu_Y(T) + S\beta + \delta U + \epsilon. \end{aligned}$$

Here, $\pi_S(T)$ is the $(G-1)\times 1$ vector of first-stage effects in experiment T (that is, the ATEs on the $G-1$ secondary metrics S of the intervention trialed in experiment T), $\mu_S(T)$ is a $(G-1)\times 1$ vector of per-experiment fixed effects on S, and $\mu_Y(T)$ is a scalar per-experiment fixed effect on Y.

Suppose that A is randomized, and that the errors are such that $E[\eta \mid S, U, A] = 0$ and $E[\epsilon \mid S, U, A] = 0$. Then the statistical estimands $\tau_S(t)$ and $\tau_Y(t)$ identify the causal parameters $\pi_S(t)$ and $\pi_S(t)\beta$. It is straightforward to check that under the data-generating process induced by this structural model,

$$\beta = \theta_1(\Lambda_K) = \theta_{2,\Psi}(\Lambda_K)$$

for any positive definite Ψ. The equality of the two statistical parameters can be understood as follows. Under the DGP induced by the above structural model, in which S fully mediates the effect of A on Y, $\tau_Y(t) = \tau_S(t)\beta$ for every $t \in [K]$. This implies that $[1, -\beta]^T$ is an eigenvector associated with eigenvalue 0, which must be the smallest as Λ_K is positive semi-definite. Therefore, the TLS estimand is also the OLS estimand.

Relationship to proxies. Now consider a new experiment $T = K+1$, in which we also randomize units equiprobably between treatment ($A = 1$) and control ($A = 0$). Denoting $S(t, a)$, $Y(t, a)$ the potential outcomes generated by the structural model (that is, the random variables obtained by setting values $T = t$ and $A = a$ and sampling U and the error terms in the above equations), the ATE on Y is related to the difference in arm-specific means on S as follows:

$$\begin{aligned} &E[Y(K+1,1) - Y(K+1,0)] \\ &= E[S(K+1,1) - S(K+1,0)]\beta \\ &= (E[S \mid T = K+1, A = 1] - E[S \mid T = K+1, A = 0])\theta_i(\Lambda_K), \end{aligned}$$

where $i = 1$ or $i = (2, \Psi)$. In words, given the data-generating distribution in K historical experiments in which we observe Y and S, and the data-generating distribution in an experiment in which we observe only the short-term outcome S, we can estimate the ATE on the long-term outcome. That is, $h(S) = \theta_i(\Lambda_K)S$ is an unbiased surrogate index for Y, meaning that ATEs on $h(S)$ equal ATEs on Y.

3.2 Linear Structural Model with INSIDE-consistent Direct Effects

We now enrich the above linear structural model with direct effects $\pi_Y(T)$ on Y that are not mediated by S:

$$\begin{aligned} S &= \mu_S(T) + \pi_S(T)A + \gamma U + \eta \\ Y &= \mu_Y(T) + \pi_Y(T)A + S\beta + \delta U + \epsilon. \end{aligned}$$

It is generally impossible to disentangle the direct ($\pi_Y(T)$) and indirect/mediated ($\pi_S(T)\beta$) effects. Still, it is possible to estimate meaningful causal parameters under the assumption of INstrument Strength Independent of Direct Effect (INSIDE) from the Mendelian randomization literature [5], which states that the first-stage effects of A on S are independent of the direct effects of A on Y. In particular, INSIDE requires that the vector $\boldsymbol{\pi}_Y = [\pi_Y(1), \ldots, \pi_Y(K)]^\top$ of direct effects is orthogonal to the columns of the matrix $\boldsymbol{\Pi}_S = [\pi_S(1)^\top, \ldots, \pi_S(K)^T]^\top$ of first-stage effects on S, that is, $\boldsymbol{\pi}_Y^T \boldsymbol{\Pi}_S = 0$.

Under the data-generating distribution induced by randomized treatment assignment and the INSIDE-consistent causal model, the covariance matrix Λ_K of true treatment effects is

$$\Lambda_K = \begin{bmatrix} \beta^T \Lambda_{SS,K}\beta & (\Lambda_{SS,K}\beta)^\top \\ \Lambda_{SS,K}\beta & \Lambda_{SS,K} \end{bmatrix} + \begin{bmatrix} \frac{\pi_Y^\top \pi_Y}{K} - \frac{(\pi_Y^\top \mathbf{1})^2}{K^2} & 0 \\ 0 & 0, \end{bmatrix}$$

with $\Lambda_{SS,K} = \frac{\Pi_S^\top \Pi_S}{K} - \frac{\mathbf{1}^\top \Pi_S \Pi_S^\top \mathbf{1}}{K^2}$.

As can be seen from the above matrix expression, $\theta_1(\Lambda_K) = \beta$ identifies the structural parameter β. However, with direct effects, it is no longer the case in general that $[1, -\beta]^\top$ is an eigenvector of Λ_K and therefore the TLS estimand $\theta_{2,\Psi}(\Lambda_K)$ diverges in general from β.

S-mediated ATE. In the presence of direct effects, it is no longer the case that ATEs on Y are given by ATEs on S times β, as S no longer mediates the effect of the treatment fully. Still, β, which coincides with the OLS estimand $\theta_1(\Lambda_K)$, has a causal interpretation under the linear structural model. Let us introduce the potential outcomes $S(t, a)$ and $Y(t, a, s)$ generated by the structural model (that is, the random variables obtained by setting $T = t, A = a, S = s$ in the above equations and independently sampling U and the error terms). It holds that the natural indirect effect on Y through S is given by

$$\begin{aligned} &E[Y(K+1, a, S(K+1,1)) - Y(K+1, a, S(K+1,0))] \\ &= E[S(K+1,1) - S(K+1,0)]\beta \\ &= (E[S \mid T = K+1, A = 1] - E[S \mid T = K+1, A = 0])\theta_1(\Lambda_K), \end{aligned}$$

for either $a \in \{0, 1\}$. That is, we can identify from the population distribution of historical experiments a proxy $h(S) = \theta_1(\Lambda_K)S$ such that in a new experiment, the ATE on $h(S)$ equals the part of the effect of the intervention on Y that is mediated by S.

3.3 Nonparametric IV Model

Relaxing the assumption of linearity, we now consider the following nonparametric IV model:

$$\begin{aligned} S &= \mu_S(T) + \pi_S(T)A + \eta \\ Y &= \mu_Y(T) + h(S) + \epsilon, \end{aligned}$$

where $E[\eta \mid A, T] = 0$ and $E[\epsilon \mid A, T] = 0$. We now show that, under some assumptions, θ_1 identifies a functional of h.

Assumption 1 (Small effects). *$\|\pi_S(t)\|_\infty \leq \epsilon$ for every $t \in [K]$ for some $\epsilon > 0$.*

We will think of ϵ as a small quantity, which is often reasonable in digital experiments.

Assumption 2 (Bounded Hessian). *h is twice-differentiable and $\left\|\nabla^2 h\right\| \leq M$ for some $M > 0$, where the norm is the nuclear norm.*

Proposition 1. Suppose that Assumptions 1-2 hold. Then,

$$\theta_1(\Lambda_K) = \left(\sum_{t=1}^{K} \pi_S(t)\pi_S(t)^\top\right)^{-1} \sum_{t=1}^{K} \pi_S(t)\pi_S(t)^\top E[\nabla h(S(t,0))] + O(M\epsilon),$$

where $S(t,0)$ is the potential outcome generated by the structural model under the control treatment in experiment t.

Proof. By definition of $\theta_1(\Lambda_K)$, the randomization of A, and from a second-order Taylor expansion,

$$\begin{aligned}
&\theta_1(\Lambda_K)\\
&= \left(\sum_{t=1}^{K} \tau_S(t)\tau_S(t)^\top\right)^{-1} \sum_{t=1}^{K} \pi_S(t) E[h(S(t,0)+\pi_S(t)) - h(S(t,0))]\\
&= \left(\sum_{t=1}^{K} \pi_S(t)\pi_S(t)^\top\right)^{-1} \sum_{t=1}^{K} \pi_S(t)\pi_S(t)^\top E[\nabla h(S(t,0))]\\
&\quad + \left(\sum_{t=1}^{K} \pi_S(t)\pi_S(t)^\top\right)^{-1} \sum_{t=1}^{K} \pi_S(t)\pi_S(t)^\top E[\nabla^2 h(\widetilde{S}(t,0))\tau_S(t)],
\end{aligned}$$

for some $\widetilde{S}(t,0)$ on the segment $[S(t,0), S(t,0)+\tau_S(t)]$. Assumptions 1 and 2 then yield that the second term above is $O(M\epsilon)$. □

In words, Proposition 1 above tells us that under the NPIV model above, the OLS statistical estimand identifies an instrument-strength-weighted average of the expected gradient of the structural function h.

4 ESTIMATING TREATMENT EFFECT COVARIANCES

In what follows, we will concatenate the variables S and Y in the vector $D = [Y, S]$ and also write $\tau(t) = [\tau_Y(t), \tau_S(t)]$.

4.1 A Naive Estimator

A naive estimator for the covariance matrix Λ_K of true treatment effects is simply the empirical covariance matrix $\hat{\Sigma}_K$ of the estimated ATEs $\widehat{\tau}_S(t)$ and $\widehat{\tau}_Y(t)$:

$$\widehat{\Sigma}_K = \frac{1}{K}\sum_{t=1}^{K} \widehat{\tau}(t)\widehat{\tau}(t)^\top - \left(\frac{1}{K}\sum_{t=1}^{K}\widehat{\tau}(t)\right)\left(\frac{1}{K}\sum_{t=1}^{K}\widehat{\tau}(t)\right)^\top,$$

with $\widehat{\tau}(t) = \frac{1}{n/2}\sum_{i=1:A_i=1,T_i=t} D_i - \frac{1}{n/2}\sum_{i=1:A_i=0,T_i=t} D_i$.

What is the target estimand of $\widehat{\Sigma}_K$? The total variance formula yields that:

$$E\widehat{\Sigma}_K = \Lambda_K + \frac{4}{n}\bar{\Omega}_K,$$

where

$$\bar{\Omega}_K = \frac{1}{K}\sum_{t=1}^{K} \frac{\Omega_{t,1}+\Omega_{t,0}}{2}$$

and $\Omega_{t,a} = \mathrm{Cov}(D \mid T = t, A = a)$ is the within-cell covariance matrix of D. In other words, $4\bar{\Omega}_K/n$ is the covariance of the unit-level sampling variance, or "noise."

In industrial (especially digital) experimentation, ATEs typically exhibit a low signal-to-noise ratio. This implies that the term $4\bar{\Omega}_K/n$ is often significant relative to Λ_K [7, 16]. As a result, under our first or second structural model, the estimate of β based on the empirical covariance matrix, $\widehat{\beta} = \widehat{\Sigma}_{SS,K}^{-1}\widehat{\Sigma}_{SY,K}$, is biased and remains inconsistent for β even as we increase the number of experiments (that is, as $K \to \infty$). Since the two-stage least squares estimator under a categorical instrument equals OLS on the group means, it holds that $\widehat{\beta}$ is 2SLS with

- dependent variable $(2A-1)Y$,
- endogenous variable $(2A-1)S$,
- instrument T.

The fact that $\widehat{\beta}$ is biased when noise is non-negligible reflects the well-established fact that 2SLS is inconsistent under weak instruments [2]. This suggests that insights from the weak instrumental variable literature could improve the estimation of Λ_K. In the next sections, we will consider three different methods inspired by weak IV estimators: (1) the Jackknife Instrumental Variable estimator (JIVE), (2) the Limited Information Maximum Likelihood (LIML) estimator, and (3) the general form of k-class estimators.

4.2 Jackknifed Covariance Matrix of Treatment Effects

We propose an estimator of Λ_K inspired by the JIVE estimator [2]. Consider the transformed vector $\widetilde{D} = 2(2A-1)D$. Observe that $E[\widetilde{D} \mid T = t] = \tau(t) = E[D \mid T = t, A = 1] - E[E \mid T = t, A = 0]$.

For any t, a, i, let

$$\widehat{\tau}(t) = \frac{1}{n}\sum_{i:T_i=t}\widetilde{D}_i \qquad \text{and} \qquad \widehat{\tau}_{-i}(t) = \frac{1}{n-1}\sum_{\substack{j\neq i\\ T_j=t}}\widetilde{D}_j$$

be the estimated ATE on D in experiment t, and its counterpart that leaves out observation i. Let

$$\widehat{\Lambda}_2(t) = \frac{1}{n}\sum_{i:T_i=t}\widehat{\tau}_{-i}(t)\widetilde{D}_i^\top$$

be the Jackknifed second-order moments matrix in experiment t, and let

$$\widehat{\Lambda}_K^{\mathrm{JK}} = \frac{1}{K}\sum_{t=1}^{K}\widehat{\Lambda}_2(t) - \left(\frac{1}{K}\sum_{t=1}^{K}\widehat{\tau}(t)\right)\left(\frac{1}{K}\sum_{t=1}^{K}\widehat{\tau}(t)\right)^\top$$

be the Jackknifed covariance matrix. The Jackknife construction ensures that only the common source of variation between units in the same experiments is captured: the variation due to the treatment assignment, as opposed to unit-level noise. It is immediate that the Jackknifed second-order moment matrix is an unbiased estimate of the second-order moment matrix of the true treatment effects. It can also readily be checked that under mild conditions the second order term has bias $O(N^{-1})$; in other words, the bias scales with the total number of units in all experiments as opposed to the number of observations per experiment. We state this formally in the following proposition:

Proposition 2. Suppose that $\max_{t,a}\|\Omega_{t,a}\| \leq M$ for some $M > 0$, where the norm is any matrix norm. Then $E[\widehat{\Lambda}_K^{\mathrm{JK}}] = \Lambda_K + O(M/N)$.

PROOF. As the second-order moment Jackknifed matrix $K^{-1}\sum_{t=1}^{K}\widehat{\Lambda}_2(t)$ is an unbiased estimate of $K^{-1}\sum_{t=1}\tau(t)\tau(t)^\top$, the bias reduces to that of the second term, which we rewrite as $N^{-1}Z_N Z_N^\top$, where $Z_N = N^{-1/2}\sum_{i=1}\widetilde{D}_i - E[\widetilde{D}_i]$. The (sequence of) random variables Z_N can be controlled under various set of conditions. For example, suppose $0 \leq m \preccurlyeq \Omega_{t,a} \preccurlyeq M < \infty$ for every t, a, and suppose that $K^{-1}\sum_t \Omega_{t,1} + \Omega_{t,0} \to \Omega$ in probability for some Ω. Then, under a Lyapunov condition, a standard central limit theorem guarantees that $Z_N \rightsquigarrow Z$ where $Z \sim \mathcal{N}(0,\Omega)$. Then $E[(Z_N Z_N^\top)] \to E[ZZ^\top)] \preccurlyeq MI_G$, which implies the result. □

As a corollary, if it admits a probability limit, the plug-in estimator $\theta_i(\widehat{\Lambda}_K^{\text{JK}})$ for either $i = 1$ or $i = (2, \Psi)$ is a consistent estimator of the parameter $\theta_i(\Lambda_K)$ as the number of experiments grows, $K \to \infty$, even as the size of each experiment remains fixed. The former is equivalent to the JIVE estimator, while the latter is equivalent to the estimator proposed in [11].

Note that the Jackknifed covariance matrix estimator does not require homoskedasticity of the unit-level noise — that is, it does not require $\Omega_{a,t} = \Omega$ for all t, a — for consistency. However, it does require us to Jackknife the unit-level data in every experiment, which can be computationally prohibitive when n and K are large. In the next two sections, we therefore consider how we can leverage homoskedasticity when it is a reasonable assumption.

4.3 Estimating Treatment Effect Covariances by Isotropizing Noise

In this section and the next, we assume a common noise covariance matrix Ω across all experiments and treatment arms, that is, $\Omega_{t,a} = \Omega$ for every t, a. We will further assume that we know Ω to a high relative precision. These are often reasonable assumptions in digital experiments: while treatment effects are small, correlations across metrics tend to be (1) stable across experiments, and (2) non-negligible, and thus easy to estimate with a high signal-to-noise ratio by leveraging populations across multiple experiments (e.g., the entire user base). Furthermore, considering metrics that are sufficiently nonredundant ensures that Ω is well-conditioned.

Under the homoskedasticity assumption, the total covariance formula yields that

$$E\widehat{\Sigma}_K = \Lambda_K + \frac{4}{n}\Omega.$$

Under known Ω, we can multiply $\widehat{\Sigma}_K$ on both sides by $\Omega^{-1/2}$ to obtain a transformed Λ_K plus isotropic noise:

$$E\Omega^{-1/2}\widehat{\Sigma}_K\Omega^{-1/2} = \Omega^{-1/2}\Lambda_K\Omega^{-1/2} + \frac{4}{n}I_G.$$

Because adding a multiple of the identity to a matrix does not affect its eigenvectors nor the rank of their corresponding eigenvalues, the smallest eigenvector of $\Omega^{-1/2}\Lambda_K\Omega^{-1/2}$ is the smallest eigenvector of $E\Omega^{-1/2}\widehat{\Sigma}_K\Omega^{-1/2}$ (where by smallest eigenvector we mean an eigenvector assoicated with the smallest eigenvalue). Denote this eigenvector by $\widetilde{\gamma}_K$. Because Λ_K and Ω both have non-negative eigenvalues, applying the transformation $\Omega^{-1/2}$ also does not affect the ordering of their eigenvectors, so we can recover the smallest eigenvector of Λ_K by applying $\Omega^{-1/2}$ to $\widetilde{\gamma}_K$. Denote by $\widehat{\gamma}_K$ the result of this procedure applied to the estimate $\widehat{\Sigma}_K$ itself (that is, rather than its unknown expectation), and let $\widehat{\theta}^{LIMLK} = -\widehat{\gamma}_{K,S}/\widehat{\gamma}_{K,Y}$. Under the existence of the appropriate probability limits, $\widehat{\theta}^{LIMLK}$ is a consistent estimate of $\theta_{2,\Omega}(\Lambda_K)$.

As our notation suggests, $\widehat{\theta}^{LIMLK}$ is equal to the LIMLK (LIML with Known noise covariance matrix) estimate with dependent variable $\widetilde{Y} = 2(2A-1)Y$, endogenous predictors $\widetilde{S} = 2(2A-1)S$, and instrument T. (This can be observed directly from the definition of LIMLK [1].)

As the smallest eigenvector of the covariance matrix of observations is the statistical target of Total Least Squares (TLS), the procedure we just described, and thus LIMLK, can be implemented by: (1) transforming the observations by applying $\Omega^{-1/2}$, (2) running TLS in the transformed space, and (3) transforming the smallest eigenvector obtained from TLS by applying $\Omega^{-1/2}$. We illustrate this procedure visually in Section 5.

A causal inference implication of this method is that, under the linear structural model presented in subsection 3.1, the structural coefficient β equals $\theta_{2,\Omega}(\Lambda_K)$, which we can estimate consistently with $\widehat{\theta}^{LIMLK}$. However, under the presence of direct effects, as mentioned in Section 3.2, $\theta_{2,\Omega}(\Lambda_K)$ no longer equals β, and therefore treatment effects on $\theta_{2,\Omega}(\Lambda_K)S$ cannot be interpreted as the part of the treatment effect of A on Y that is mediated by S.

4.4 Estimating Treatment Effect Covariances by Subtracting Noise

In the previous subsection, we leveraged only the direction of the known Ω but not its scale, allowing us to make a connection to LIMLK. (Note that the above procedure would give the same result if we used $\rho\Omega$ for $\rho > 0$ instead of Ω.) Under known Ω, there is a perhaps more straightforward estimation procedure for Λ_K: subtract $(4/n)\Omega$ from $\widehat{\Sigma}_K$. Formally, letting $\widehat{\Lambda}_K^{\text{TC}} = \widehat{\Sigma}_K - (4/n)\Omega$ (where TC stands for Total Covariance), we have that

$$E\widehat{\Lambda}_K^{\text{TC}} = \Lambda_K,$$

and therefore, under the existence of the appropriate probability limits, the plug-in estimator $\theta_i(\widehat{\Lambda}_K^{\text{TC}})$, for either $i = 1$ or $i = (2, \Psi)$, provides a consistent estimator for $\theta_i(\Lambda_K)$. In particular, when either S fully mediates the effect of A on Y (no direct effects) or when direct effects follow the INSIDE assumption, we can consistently estimate the structural parameter β with $\theta_1(\widehat{\Lambda}_K^{\text{TC}})$.

Connection with IV-estimators. Defining the matrix of centered observations $\widetilde{D}^0 = [\widetilde{Y}^0, \widetilde{S}^0]$ as $\widetilde{D}^0 = D(I - N^{-1}\mathbf{1}\mathbf{1}^\top)\widetilde{D}$, and using the empirical within-experiment covariance for Ω, we can readily check that

$$\theta_1(\widehat{\Lambda}_K^{\text{TC}}) = \frac{(\widetilde{S}^0)^\top(I - (1+4/n)M_T)\widetilde{Y}^0}{(\widetilde{S}^0)^\top(I - (1+4/n)M_T)\widetilde{S}^0}, \tag{1}$$

with $M_T = I - P_T$, $P_T = \widetilde{T}(\widetilde{T}^\top\widetilde{T})^{-1}\widetilde{T}^\top$, and $\widetilde{T}$ is the $N \times K$ matrix of one-hot encodings of experiment membership T (that is $\widetilde{T}_{i,t} = \mathbf{1}\{T_i = t\}$). One might recognize from (1) that $\theta_1(\widehat{\Lambda}_K^{\text{TC}})$ is a so-called k-class IV estimator, with $k = 1 + 4/n$.

5 SIMULATION STUDY

To provide insight into our statistical setup and the performance of our estimators, we conduct a simulation study. The parameters of our simulations are chosen to reflect aspects of real-world data. The unit-level noise, denoted by Ω in our statistical setup, is typically large relative to the variance-covariance matrix of treatment effects, denoted by Λ_K, and is also potentially anti-correlated. For example, clicks may be positively correlated with conversions outside of any experiment, but treatments that increase "click-baitiness" can reduce conversions.

Following this example, the top panel of Figure 1 depicts a situation where treatment effects on a proxy metric S are negatively correlated with treatment effects on a primary objective Y. Throughout the figure, the white arrow points in the direction of the covariance in true treatment effects. In the middle panel, we plot the unit-level noise, which is positively correlated between S and Y, represented by the black line. Lastly, in the bottom panel, we plot the sampling distribution of treatment effect estimates with a fixed sample size.

As bottom panel of Figure 1 shows, the unit-level noise can overwhelm the treatment effect covariance when either the treatment effect covariance or the sample size is relatively small. As a result, naively estimating $\theta_1(\Lambda_K)$ using the covariance matrix of the estimated treatment effects $\theta_1(\widehat{\Sigma}_K)$ will be biased in the direction of $\theta_1(\Omega)$. In the absence of unit-level covariance (i.e., $\Omega = \omega I$), this bias is "merely" attenuation bias that preserves the direction of the relationship but biases estimates towards zero. In the presence of unit-level covariance, the estimated covariance can be arbitrarily biased, and this bias is worse for small experiments [7].

5.1 LIMLK as Total Least Squares

We now consider the LIMLK estimator as an alternative to the naive estimator based on the covariance matrix of estimated treatment effects. As described in Section 4.3, LIMLK is equivalent to performing Total Least Squares (TLS) after applying the linear transformation $\Omega^{-1/2}$ to the scatterplot of estimated treatment effects. Intuitively, this is because applying this transformation on both sides to $\widehat{\Sigma}_K$ yields, in expectation, the transformed true covariance matrix $\Omega^{-1/2}\Lambda_K\Omega^{-1/2}$ plus isotropic noise — a classic error-in-variables setup. TLS is an effective method for addressing error-in-variables as isotropic noise does not change the eigenvectors of $\widehat{\Sigma}_K$.

We illustrate the effects of transforming $\widehat{\Sigma}_K$ by $\Omega^{-1/2}$ on least squares estimators in Figure 2. In both panels, the blue arrow is the statistical parameter $\theta_1(\Omega^{-1/2}\Lambda_K\Omega^{-1/2})$, which we call the **OLS target**. This is the OLS on the transformed scatterplot of true treatment effects. In contrast, the orange line is the statistical parameter $\theta_1(\Omega^{-1/2}E\widehat{\Sigma}_K\Omega^{-1/2})$, which we call the **OLS plim**. In words, this is the OLS on the transformed scatterplot of the estimated treatment effects as $K \to \infty$. The white arrow and gray line in Figure 2 are the analogous **TLS target**, $\theta_{2,\Omega}(\Lambda_K)$, and **TLS plim**, $\theta_{2,\Omega}(E\widehat{\Sigma}_K)$, respectively. To illustrate the effect of sample size, the upper panel sets $n = 100$, while the bottom panel sets $n = 10000$.

Because transforming the scatterplot of estimated treatment effects isotropizes but does not eliminate the noise, the OLS plim of the transformed scatterplot will suffer from attenuation bias relative to the OLS target, with the magnitude of this bias decreasing in n (evidenced by the convergence of the estimand to the target as $n \to \infty$). However, isotropic noise does not change the eigenvectors of the covariance matrix, so increasing n does not affect the TLS plim, which always aligns with the TLS target. This motivates the use of TLS on the transformed scatterplot, which we show in Section 4.3 is equivalent to the LIMLK estimator.

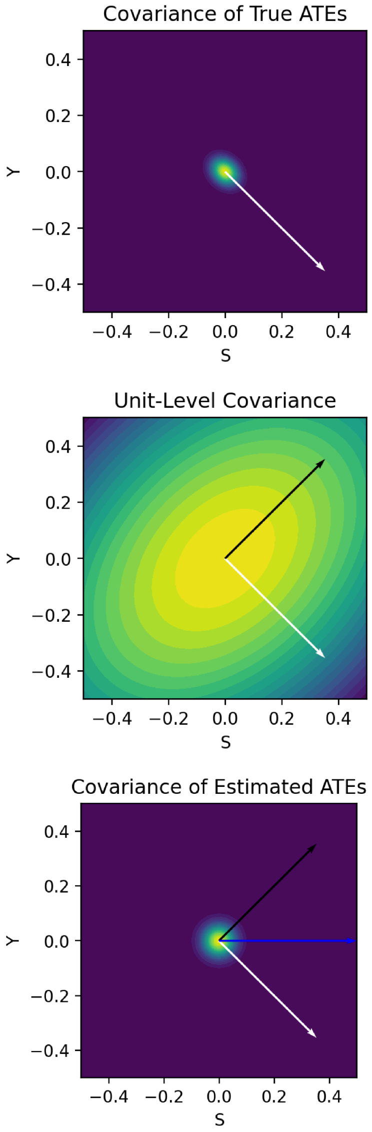

Figure 1: How Measurement Error Distorts Treatment Effect Covariances. This figure shows a hypothetical treatment effect covariance matrix, a unit-level sampling covariance matrix ("noise"), and the covariance of matrix of *estimated* treatment effects, which is a weighted combination of these.

5.2 Comparing the Naive, LIMLK, and TC Estimators

We now present the results of a simulation study of the empirical properties of our estimators. The estimators evaluated are the naive OLS performed on the empirical covariance matrix of estimated effects, LIMLK, and the Total Covariance (TC) estimator, which

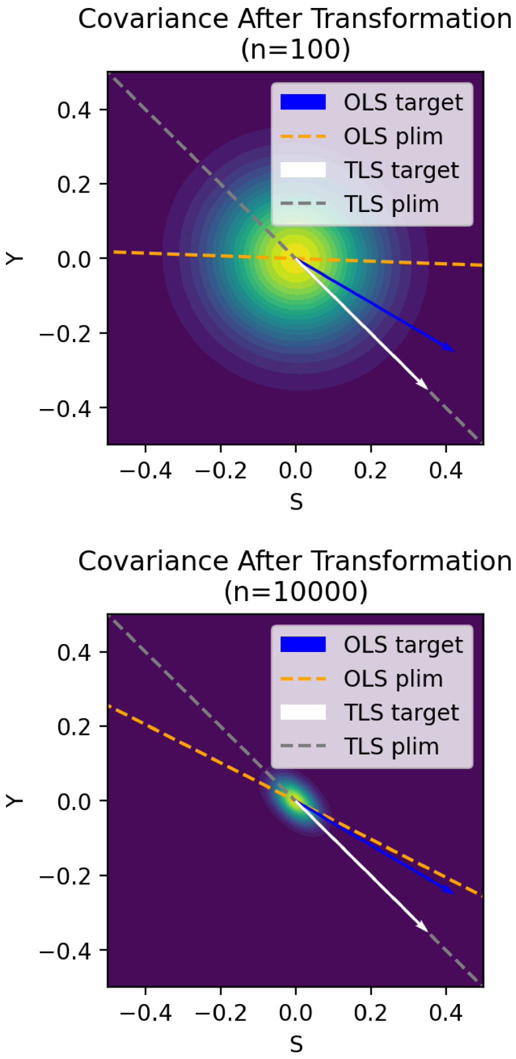

Figure 2: Effect of Transforming Estimated Treatment Effects by $\Omega^{-1/2}$. Transforming the covariance of estimated treatment effects renders the measurement error isotropic, so the TLS plim on this scatterplot coincides with the TLS target, whereas the OLS plim suffers from attenuation bias.

performs OLS on the covariance matrix obtained by subtracting $(4/n)\Omega$ from $\widehat{\Sigma}_K$.

To illustrate the effects of unit-level noise on our estimators, we simulate two surrogates S_1 and S_2 with corresponding structural parameters β_1 and β_2. S_1 is highly correlated with the primary outcome Y in terms of the noise covariance matrix Ω, while S_2 is uncorrelated with Y. As a result, we expect the naive estimator to perform relatively badly for S_1 (as it ignores the unit-level covariance) compared to S_2 and the LIMLK and TC estimators. To assess the robustness of our estimators to direct effects, we simulate data according to the linear structural models in Subsections 3.1 (no direct effects) and 3.2 (direct effects under INSIDE).[3]

As Figure 3 shows, the naive estimator is heavily biased. LIMLK and TC are substantially much more accurate than the naive estimator. The relative accuracy of LIMLK and TC is more apparent for β_1, which has the greater distortion from the unit-level covariance. Without direct effects, LIMLK is more efficient compared to TC, with a considerably smaller standard error at smaller values of n.

However, LIMLK is highly sensitive to the assumption of no direct effects. As Figure 4 shows, when we introduce direct effects, LIMLK is clearly inconsistent for β. In contrast, TC is robust and remains consistent for β. However, note that, regardless of estimator, the *interpretation* of β is sensitive to the presence of direct effects. If S fully mediates the effect of A on Y, $h(S) = \beta S$ is a surrogate index, and the ATE of a new treatment on Y is equal to the ATE on $h(S)$. If S does not fully mediate the effect of A on Y, and direct effects follow the INSIDE assumption, then β can be interpreted as the portion of the treatment effect of Y that is mediated by S.

6 LINEAR MODELS OF TREATMENT EFFECT COVARIANCES AT NETFLIX

This section discusses the real-world application of our methods to construct linear proxy metric indices for experimentation at Netflix. Netflix has a sophisticated experimentation platform that runs thousands of experiments on millions of experimental units [10]. Still, because the signal-to-noise ratio of each experiment is small, measurement error poses challenges. This is shown in the top panel of Figure 5, which plots the correlation between *estimated* treatment effects on one Short-term Metric and one Long-term Metric across 96 arbitrarily-chosen treatment-control comparisons. The Short- and Long-term Metrics are highly correlated across experimental units, introducing correlated measurement error that exaggerates the correlation across treatment effects. To demonstrate the effect of experiment size on this bias, we show the scatterplot of estimated treatment effects after downsampling each experiment to one million units (white circles) and 15 million units (gray circles). The slope of the OLS in the former scatterplot is about 50% larger than the equivalent slope in the latter scatterplot. Note that we have hidden the x- and y-axis tick marks in this figure for confidentiality reasons.

In the bottom panel of Figure 5, we plot the absolute bias of the observed covariance between the Short- and Long-term Metrics at various subsample sizes, using as ground truth the TC-corrected covariance across non-downsampled experiments. We also plot the absolute bias of the TC-corrected covariance for comparison. The median absolute bias reduction from applying our TC estimator is substantial — approximately 63%.

In addition to these empirical benefits, our method is also well-suited to the decentralized and rapidly-evolving practice of experimentation at Netflix. Netflix runs thousands of experiments per year on many diverse parts of the business. Each area of experimentation is staffed by independent data science and engineering teams. While every team ultimately aims to lift the same north star metrics

[3] Code to replicate our simulations and all figures except those in Section 6 can be found at https://github.com/winston-chou/linear-proxy-metrics.

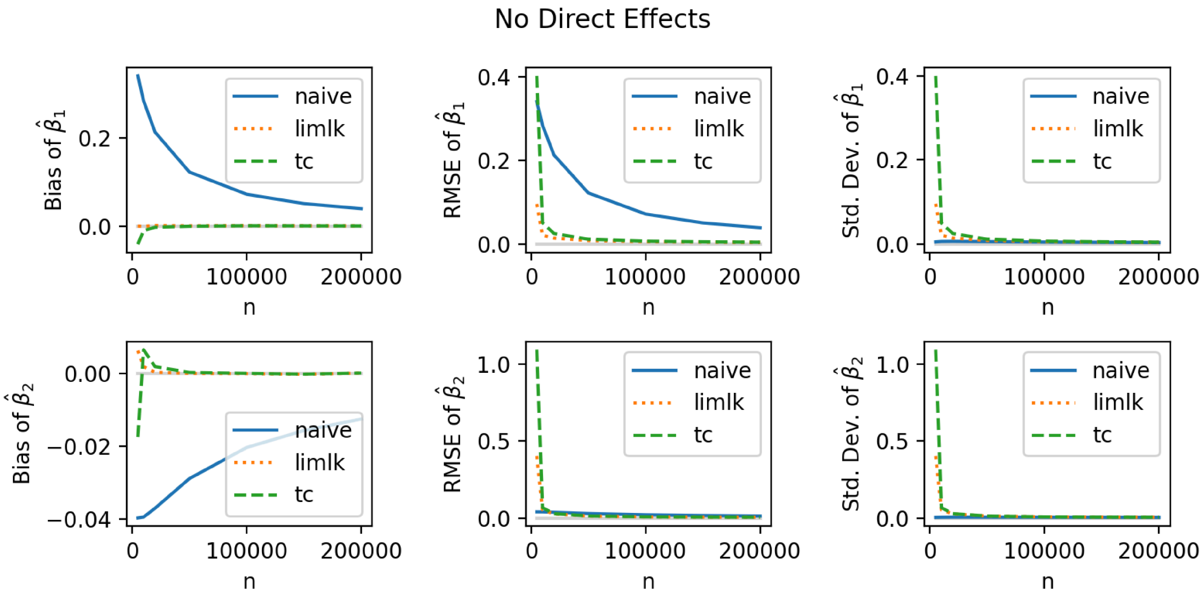

Figure 3: Comparison of Naive, LIML, and Total Covariance Estimators Without Direct Effects. When there are no direct effects, such that the secondary outcomes fully mediate the effect of the intervention on the primary outcome, our LIMLK (`limlk`) and Total Covariance (`tc`) estimators are substantially less biased than the naive regression of the estimated ATEs on the primary outcome on the estimated ATEs on the secondary outcomes (`naive`). LIMLK has lower variance than TC, which is most apparent at small sample sizes.

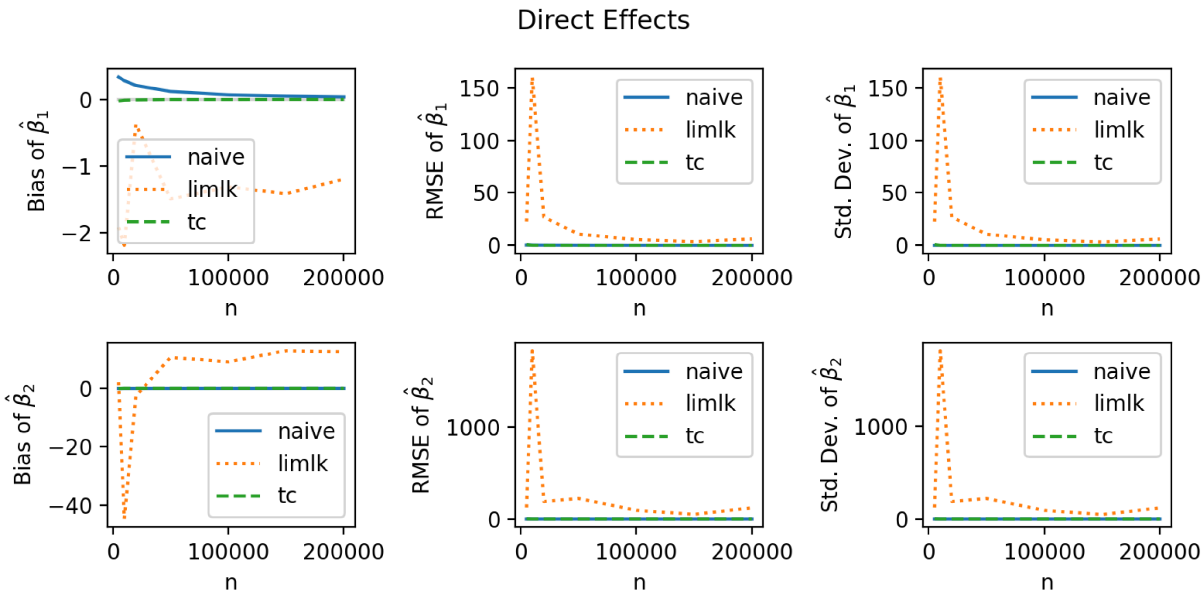

Figure 4: Comparison of Naive, LIML, and Total Covariance Estimators With Direct Effects. When the intervention directly affects the primary outcome, LIMLK can be extremely biased. In contrast, TC is robust to direct effects, although the causal interpretation of its estimand changes if there are direct effects. The naive estimator is biased regardless of whether there are direct effects or not.

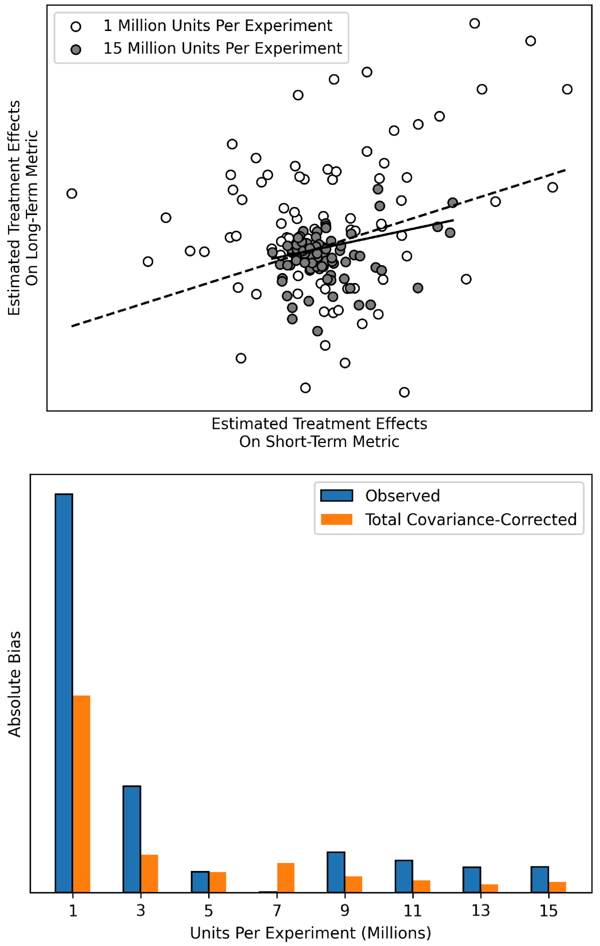

Figure 5: How Our Method Improves Treatment Effect Covariance Estimation at Netflix. This figure shows how correlated measurement error exaggerates the correlation between *estimated* treatment effects, especially at smaller sample sizes (top), and how our TC estimator reduces this bias by subtracting a scaled estimate of the measurement error (bottom).

(e.g., long-term revenue), each has also developed secondary metrics that are more sensitive and practical to measure in short-term A/B tests (e.g., user engagement or latency). To complicate matters further, teams are constantly innovating on these secondary metrics to find the right balance of sensitivity and impact on the north star metrics.

In this decentralized environment, linear models of relationships between treatment effects are a highly useful tool for coordinating these efforts and aligning them towards the same proxies for the north star:

(1) **Managing metric tradeoffs.** Because experiments in one area can affect metrics in another area, there is a need to measure all secondary metrics in all tests, but also to understand the relative impact of these metrics on the north star. This is so we can inform decision-making when one metric trades off against another metric.
(2) **Informing metrics innovation.** To minimize wasted effort on metric development, it is also important to understand how metrics correlate with the north star "net of" existing metrics.
(3) **Enabling teams to work independently.** Lastly, teams need simple tools in order to iterate on their own metrics. Teams may come up with dozens of variations of secondary metrics, and slow, complicated tools for evaluating these variations are unlikely to be adopted.

Given these needs and the availability of statistics from historical experiments, it has been common for teams to fit linear models to estimated treatment effects to identify promising surrogates and estimate their weights in a linear proxy metric index. While data scientists are aware that these models can be highly biased, linear models have a convenience and interpretability that is difficult to replace. Our TC estimator provides a simple way to consistently and robustly estimate the true treatment effect covariance matrix, which supports the construction of these linear proxies, and is actively used to develop proxy metrics at Netflix.

7 CONCLUSION

In this paper, we discuss the construction of proxy metrics via meta-analysis of many experiments. A useful parameter in this setting is the covariance matrix of treatment effects across metrics in the population of experiments. While estimating this parameter is computationally convenient, it presents challenges, especially when treatment effects are small relative to unit-level noise. We present estimators for overcoming these challenges and show how their estimands relate to structural parameters under different causal models.

Throughout, we have assumed that the unit-level covariance matrix is correctly specified and constant across experiments. This assumption is often reasonable given that it can be estimated on vastly more units than are in any individual experiment (e.g., across the entire user base). In cases where it is not reasonable, users can also compute Jackknife covariance matrices. That being said, this is a computationally expensive approach, and future work can explore more scalable ways of relaxing this assumption.

We show that LIMLK is efficient but inconsistent under direct effects while TC is consistent under direct and indirect effects. This suggests that it is possible to construct a statistical test for direct effects by comparing the LIMLK and TC residuals. We also leave this as a direction for future research.

Beyond the construction of proxy metrics, there are other opportunities to leverage short-term observations for inference on long-term outcomes. For example, [3, 6, 13] consider how short-term observations from one randomized experiment can remove confounding in an observational study with long-term observations, and [14] consider efficiency gains from including units with no long-term observations. Considering the use of many, albeit weak, experiments as instruments in these settings may uncover new opportunities to target long-term outcomes under weaker assumptions.

REFERENCES

[1] Anderson, T., Kunitomo, N., Matsushita, Y., et al. The limited information maximum likelihood estimator as an angle. *CIRJE No. CIRJE-F-619, CIRJE, Faculty of Economics, University of Tokyo* (2009).

[2] Angrist, J. D., Imbens, G. W., and Krueger, A. B. Jackknife instrumental variables estimation. *Journal of Applied Econometrics 14*, 1 (1999), 57–67.

[3] Athey, S., Chetty, R., and Imbens, G. Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676* (2020).

[4] Athey, S., Chetty, R., Imbens, G., and Kang, H. Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index. *arXiv preprint arXiv:1603.09326* (2016).

[5] Burgess, S., and Thompson, S. G. Interpreting findings from mendelian randomization using the mr-egger method. *European journal of epidemiology 32* (2017), 377–389.

[6] Chen, J., and Ritzwoller, D. M. Semiparametric estimation of long-term treatment effects. *Journal of Econometrics 237*, 2 (2023), 105545.

[7] Cunningham, T., and Kim, J. Interpreting experiments with multiple outcomes.

[8] Deng, A., Xu, Y., Kohavi, R., and Walker, T. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining* (2013), pp. 123–132.

[9] Elliott, M. R., Conlon, A. S., Li, Y., Kaciroti, N., and Taylor, J. M. Surrogacy marker paradox measures in meta-analytic settings. *Biostatistics 16*, 2 (2015), 400–412.

[10] Forsell, E., Beckley, J., Ejdemyr, S., Hannan, V., Rhines, A., Tingley, M., Wardrop, M., and Wong, J. Success stories from a democratized experimentation platform. *arXiv preprint arXiv:2012.10403* (2020).

[11] Hahn, J., Hausman, J., and Kuersteiner, G. Estimation with weak instruments: Accuracy of higher-order bias and mse approximations. *The Econometrics Journal 7*, 1 (2004), 272–306.

[12] Hohnhold, H., O'Brien, D., and Tang, D. Focusing on the long-term: It's good for users and business. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), pp. 1849–1858.

[13] Imbens, G., Kallus, N., Mao, X., and Wang, Y. Long-term causal inference under persistent confounding via data combination. *arXiv preprint arXiv:2202.07234* (2022).

[14] Kallus, N., and Mao, X. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv preprint arXiv:2003.12408* (2020).

[15] Mikusheva, A., and Sun, L. Inference with many weak instruments. *The Review of Economic Studies 89*, 5 (2022), 2663–2686.

[16] Peysakhovich, A., and Eckles, D. Learning causal effects from many randomized experiments using regularized instrumental variables. In *Proceedings of the 2018 World Wide Web Conference* (2018), pp. 699–707.

[17] Prentice, R. L. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine 8*, 4 (1989), 431–440.

[18] Tripuraneni, N., Richardson, L., D'Amour, A., Soriano, J., and Yadlowsky, S. Choosing a proxy metric from past experiments. *arXiv preprint arXiv:2309.07893* (2023).

[19] Wong, J., Lewis, R., and Wardrop, M. Efficient computation of linear model treatment effects in an experimentation platform. *arXiv preprint arXiv:1910.01305* (2019).

[20] Xie, H., and Aurisset, J. Improving the sensitivity of online controlled experiments: Case studies at netflix. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 645–654.