Interpreting Experiments with Multiple Outcomes

Tom Cunningham[*]and Josh Kim[†]

September 17, 2022

Abstract

When we observe the effect of an experiment on multiple outcomes, the interpretation will be sensitive to how those outcomes covary across units. We derive a number of results in a Gaussian model of multi-outcome experimentation - (1) the observed effect on one outcome will be a negative signal about the true effect on another outcome, under conditions that are likely to hold for many experiments; (2) in some cases, the inferred treatment-effect on an outcome can be *decreasing* in its own observed treatment-effect; (3) naive approaches to metric "surrogacy," when one metric is used to predict another, will be biased in the direction of the unit-level covariance, and naive causal estimates will suffer attenuation bias; (4) composite metrics, i.e. weighted averages of multiple outcomes, will often be shrunk by more than their components. Finally we show how to combine multivariate shrinkage with network effects and dynamic effects to yield a single matrix which maps outcomes of an experiment into the best estimate of the long-run aggregate impact of a policy.

Introduction

Suppose that you run an experiment with N units assigned to treatment and control groups, and define y_1 and y_2 as the observed treatment effects for outcomes 1 and 2. We can decompose the observed outcomes into treatment-effects and noise:

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}}_{\text{outcomes}} = \underbrace{\begin{pmatrix} t_1 \\ t_2 \end{pmatrix}}_{\text{treatment effects}} + \underbrace{\begin{pmatrix} e_1 \\ e_2 \end{pmatrix}}_{\text{noise}}$$

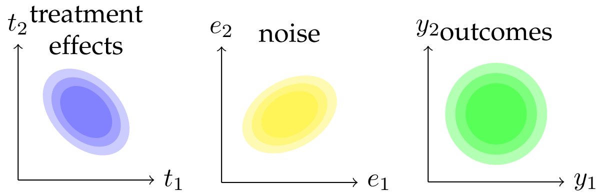
In this paper we will be particularly interested in the covariances of the treatment effects, and of the noise.

The noise terms e_1 and e_2 represent the sampling error, and therefore will have variances and covariances corresponding to the variances and covariances of the individual units, multiplied by a factor of $\frac{2}{N}$.[1].

The variances and covariances of t_1 and t_2 represent the experimenter's priors, and so are often difficult to quantify. If we are willing to identify priors with some set of previously-run experiments, i.e. an "empirical-Bayes" technique, we can recover them from the data using this relationship between covariance matrices:

$$\Sigma_y = \Sigma_t + \frac{2}{N}\Sigma_u,$$

where Σ_u is the unit-level covariance matrix. The following graph illustrates a case with negatively-correlated treatment effects, positively correlated noise, and uncorrelated outcomes.



If we assume that everything has a normal distribution, we have a crisp expression for how the posterior expectations depend on the observed outcomes. For an arbitrary number of outcomes we can write this as:

$$\mathbb{E}[t|y] = \mu_t + \Sigma_t(\Sigma_t + \frac{1}{N}\Sigma_u)^{-1}(y - \mu_t).$$

With just two outcomes it becomes:

$$\begin{aligned}\mathbb{E}[t_1|y_1,y_2] = \mu_1 + \\ |\Sigma_y|^{-1}\Big((\sigma_{t1}^2(\sigma_{t2}^2 + \sigma_{e2}^2) - \gamma_t(\gamma_t + \gamma_e))(y_1 - \mu_1) \\ + (\gamma_t\sigma_{e1}^2 - \sigma_{t1}^2\gamma_e)(y_2 - \mu_2)\Big)\end{aligned}$$

Where $\gamma_t = \text{cov}(t_1, t_2)$ and $\gamma_e = \text{cov}(e_1, e_2)$. The result can be visualized with a vector field showing

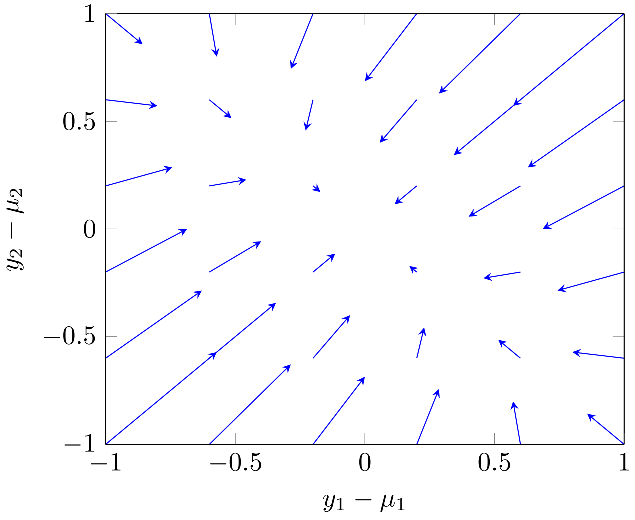
[*] Facebook, Core Data Science Economics.

[†] Stanford University, Department of Economics.

[1] Throughout we assume that the treatments change only the mean, not the variance, of outcomes.

the mapping from a pair of observed outcomes, $\binom{y_1}{y_2}$, into a pair of inferred outcomes, $E[\begin{smallmatrix}t_1\\t_2\end{smallmatrix}|\begin{smallmatrix}y_1\\y_2\end{smallmatrix}]$.

Arrows representing $E[t_1|y_1, y_2] - y_1$ and $E[t_2|y_1, y_2] - y_2$.



σ^2_{t1}=3, σ^2_{t2}=3, γ_t=0, σ^2_{e1}=1, σ^2_{e2}=1, γ_e=.8.

Interpreting multi-outcome experiments.

We can make a few broad observations, based on the equation above.

First, if there is no covariance either across treatment-effects ($\gamma_t = 0$), or across units ($\gamma_e = 0$), then the expression reduces to univariate shrinkage, i.e.:

$$\mathbb{E}[t_1|y_1, y_2] = \mathbb{E}[t_1|y_1] = \mu_1 + \frac{\sigma^2_{t1}}{\sigma^2_{t1} + \sigma^2_{e1}}(y_1 - \mu_1).$$

Some evidence on the importance of the covariance matrices comes from Coey and Cunningham (2019), which finds that multivariate shrinkage of experiment results significantly outperforms univariate shrinkage, but was not able to establish how much of this was due to cross-outcome relationships in the treatment-effects vs the noise.

Second, if there is only covariance across treatment-effects, and that covariance is positive, ($\gamma_t > 0, \gamma_e = 0$), then each observed outcome is "good news" about the treatment-effect on the other outcome, i.e.:

$$\frac{dE[t_1|y_1, y_2]}{dy_2} > 0, \frac{dE[t_2|y_1, y_2]}{dy_1} > 0.$$

Intuitively – if you expect your experiment to shift both outcomes in the same direction, then seeing a positive effect on one outcome will positively reinforce your belief in the effect on the other outcome.

Good News is Bad News (weak version).

When the relative covariance of noise is stronger than the relative covariance of the treatment-effects, then we will find that a higher-than-expected outcome on outcome 2 will be a *negative* signal about the treatment-effect on outcome 1. Formally:

$$\frac{dE[t_1|y_1, y_2]}{dy_2} < 0 \iff \frac{\gamma_t}{\sigma^2_{t1}} < \frac{\gamma_e}{\sigma^2_{e1}}$$

In many contexts we believe that this condition is likely to hold. Most desirable outcomes tend to have positive covariance across experimentation units ($\gamma_e > 0$), e.g. among people wealth, physical health, mental health, and education all tend to covary positively. Among users of online services, those who are more active on one dimension tend to be more active on all others. On the other hand, treatments tested in experiments often find trade-offs with other outcomes ($\gamma_t < 0$): e.g. promoting one part of a product tends to cannibalize time spent on other parts.

Example 1. *A school runs an experiment with a new English textbook, and they find an increase in English test-scores. Suppose they also find an increase in Maths test-scores: this would be bad news about the true impact on English ability if (a) there is strong positive correlation among students between English and Maths test-scores, (b) there is not a strong prior reason to believe that there would be a positive spillover on Maths test-scores.*

Good News is Bad News (strong version).

In some cases an even stronger result holds: y_1 is bad news about t_1 itself:

$$\frac{dE[t_1|y_1, y_2]}{dy_1} < 0 \iff 1 < \frac{\gamma_t}{\sigma^2_{t1}} \frac{\gamma_t + \gamma_e}{\sigma^2_{t2} + \sigma^2_{e2}}.$$

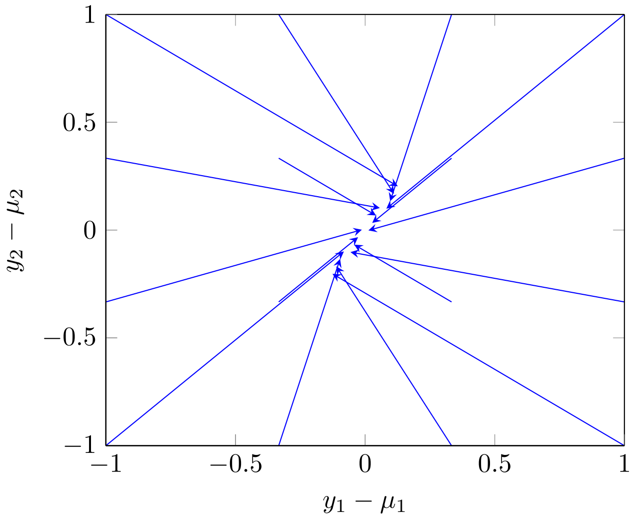
The condition requires there be non-zero covariance in *both* the treatment-effects and the unit-level outcomes (i.e., both $\gamma_t \neq 0$ and $\gamma_e \neq 0$).[2] We also can see that the two covariances must have the same sign – either both positive or both negative. Intuitively, this will occur when treatment-effects are closely correlated, such that y_2 is a relatively better signal for t_1 than y_1 itself, and so y_1 instead becomes a signal for the correlated noise.

[2] If $\gamma_e = 0$ the condition will never hold, because we know that $\frac{\gamma_t}{\sigma^2_{t1}\sigma^2_{t2}}$ represents the correlation in treatment effects, which is bounded between 0 and 1.

Example 2. *You run an experiment where you randomly select married households and provide the wife with vocational training and education, and measure the impact on both the wife's income, and the overall household's income. You find that average annual female labor income increased by $1,000. However, you also find that average annual household income increased by $3,000, implying the average male income in the treated group is $2,000 higher than in the control group. If you believe that (a) the spillover effect from wife's income to husband's income is likely to be small or negative, and (b) the correlation between wife's and husband's income is strong, then this evidence should cause you to decrease your estimate of the true impact on both the wife's income and the total income.*

The following vector-field illustrates the strong version of the "good news is bad news" effect: it can be seen that, for a given value of y_2, progressively higher values of y_1 are mapped into relatively lower values of t_1.

Arrows representing $E[t_1|y_1, y_2] - y_1$ and $E[t_2|y_1, y_2] - y_2$.



σ^2_{t1}=1, σ^2_{t2}=1, γ_t=.9,
σ^2_{e1}=6, σ^2_{e2}=2, γ_e=3.

Surrogate Metrics

When we wish to estimate effects on a noisy outcome, investigators often try to find a "surrogate" outcome that is a good predictor of the primary outcome but more precisely measured.[3] Sometimes surrogacy relationships are estimated by a simple regression across experiment outcomes. However this will give a biased estimate because the coefficient will pick up the covariance between units, as well as covariance between experiments:

$$\hat{\beta} = \frac{dE[y_2|y_1]}{dy_1} = \frac{\text{cov}(y_1, y_2)}{\text{var}(y_1)} = \frac{\gamma_t + \gamma_e}{\sigma^2_{t1} + \sigma^2_{e1}}.$$

Figure 1 plots a set of AA tests with no true treatment effect on either outcome. Despite the lack of any relationship between treatment-effects, we find a significant relationship between outcomes, due to covariance among units.

However the quantity of interest for a surrogate variable is instead:

$$\frac{dE[t_2|y_1]}{dy_1} = \frac{\text{cov}(t_2, y_1)}{\text{var}(y_1)} = \frac{\gamma_t}{\sigma^2_{t1} + \sigma^2_{e1}}.$$

To calculate the correct quantity we can estimate the covariance across treatment-effects by taking the difference between observed covariances across experiments and that expected from AA-tests: $\hat{\gamma}_t = \gamma_y - \frac{1}{N}\gamma_u$.

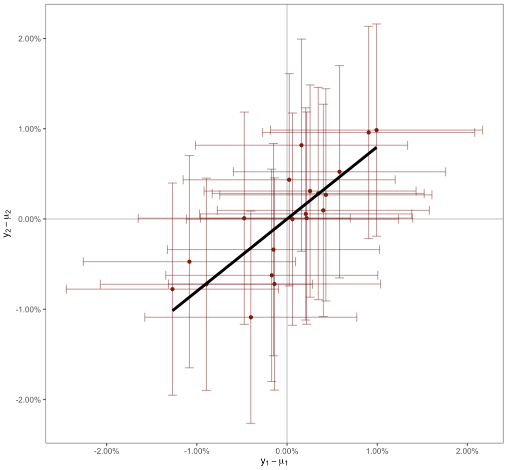


Figure 1: A simulated scatter-plot showing 20 experiments, with N=1,000,000, $\sigma^2_{e1} = \sigma^2_{e2} = 1$, with correlation 0.8. The experiments are all AA-tests, i.e. there are no true treatment effects, yet a regression of y_2 on y_1 will consistently yield statistically-significant coefficients of around 0.8.

The bias described will decline with sample-size N, but is invariant to the number of experiments. Additionally, if there is no true treatment-effect on outcome 1 ($\sigma^2_{t1} = 0$), as in the AA-tests illustrated in Figure 1, then the estimated $\hat{\beta}$ will always equal $\frac{\gamma_e}{\sigma^2_e}$, even as the sample-size goes to infinity.

[3] See Athey et al. (2016) for more discussion.

Causal Effects

In other circumstances we wish to understand the relationship between treatment-effects, i.e.:[4]

$$\frac{dE[t_2|t_1]}{dt_1} = \frac{\gamma_t}{\sigma_{t1}^2}.$$

Again this does not correspond to the quantity recovered by a regression, $\hat{\beta}$ above. There are now two biases: (1) an "activity" bias due to γ_e in the numerator of $\hat{\beta}$; (2) an attenuation bias due to σ_{e1}^2 in the denominator of $\hat{\beta}$.

Example 3. *You run a series of A/B experiments meant to improve music recommendations and are interested in how these experiments impact time-spent on music, and potential cannibalization of time-spent on podcasts. Estimating a cannibalization rate by regressing* `podcast-time-spent` *on* `music-time-spent`*, across experiments, would lead to underestimating the true rates of cannibalization for two reasons: (1) positive unit-level covariance between the two outcomes causing positive correlation between the two outcomes, (2) noise in the estimate of the treatment-effects on* `music-time-spent`*, causing attenuation towards zero.*

Composite Metrics

We sometimes want to estimate the impact of an experiment on a "composite" outcome, a linear combination of n outcomes with weights $w = (w_1, \ldots, w_n)$. It is useful to calculate the *signal-to-noise ratio* (SNR) of the composite metric, $\bar{y}$:

$$\text{SNR}_{\bar{y}} = \frac{Var[w't]}{Var[w'e]} = \frac{w'\Sigma_t w}{w'\Sigma_e w}.$$

The signal-noise ratio is a useful statistic to track because an outcome with a higher SNR will have (a) a lower shrinkage factor, i.e. the posterior $E[t|y]$ will be relatively closer to the observed y, and (b) will have a higher fraction of experiments that are statistically significantly different from zero:[5]

$$\begin{aligned}\text{shrinkage factor} =& 1 - \frac{E[t|y]-\mu}{y-\mu} = \frac{1}{1+SNR},\\ \text{fraction significant} =& P(|y-\mu| > 1.96\sigma_e)\\ =& 2\Big(1 - \Phi(\frac{1.96}{1+SNR})\Big).\end{aligned}$$

Where Φ is the CDF of a standard Normal distribution.

We can make two observations about the signal-noise ratio of a composite metric:

1. If all covariance terms are zero, across treatments and units, then the SNR of $\bar{y}$ will be a weighted average of the SNR of each of the components:

   $$\text{SNR}_{\bar{y}} = \frac{\sum w_i^2 \sigma_{ti}^2}{\sum w_i^2 \sigma_{ei}^2} = \sum_i \frac{w_i^2 \sigma_{ei}^2}{\sum_j w_j^2 \sigma_{ej}^2} \text{SNR}_{y_i}.$$

   This implies that adding a new component to a composite metric will increase its SNR (and so increase the fraction of statistically-significant experiments) if and only if the new component has a higher SNR than the existing composite metric.

2. If the outcomes have positive covariance across units, but zero covariance across treatments, the composite's signal-noise ratio will be *below* the weighted-average SNR. This can be seen in the first equation for $\text{SNR}_{\bar{y}}$: the covariance terms in the error will show up in the denominator, causing the SNR to decline.

These observations are consistent with the general observation that tech companies often struggle when dealing with metrics that are designed to reflect the full business impact of an experiment: if they add all outcomes of interest into the composite metric, the composite will be relatively noisy for two reasons: (1) because some components are noisy, and (2) because of positive covariance in the noise components.

An optimally shrunk composite metric will take into account the noise. If there is no covariance then we have:

$$\mathbb{E}[t|y_1, \ldots, y_n] = \sum_{i=1}^{n} \frac{\text{SNR}_i}{1+\text{SNR}_i} w_i y_i.$$

This estimate of the composite will appropriately shrink noisy components, and so there is no longer a penalty for adding additional components to the composite. If there is covariance then the full expression will be:

$$\mathbb{E}[\bar{t}|y] = w'\mathbb{E}[t|y].$$

Network and Dynamic Effects

Given an observed experimental outcome, y, the relevant policy question is typically the aggregate long-run effect on outcomes, and there are three important considerations: (1) adjusting for experimental noise,

[4] One example would be when we have reason to believe that t_1 causes t_2, and we wish to use experiments as instruments to estimate the causal relationship, see Peysakhovich and Eckles (2018) for more on this.

[5] Under the assumption that t is mean-zero.

as we have discussed in this paper, (2) adjusting for network-effects, and (3) adjusting for dynamic effects. Here we briefly show how all three can be expressed in a very simplifed model. Suppose the behaviour of unit i at time $t+1$ depends on (i) some user-specific constant term, a_i, (ii) their own prior behaviour, $x_{t,i}$, and (iii) the global average $\bar{x}_t = \frac{1}{n}\sum x_{i,t}$, as:

$$x_{i,t+1} = a_i + Bx_{i,t} + C\bar{x}_t.$$

Solving for equilibrium, setting $x_{i,t+1} = x_{i,t}$ we get:

$$\bar{x} = (I - B - C)^{-1}\bar{a}.$$

Where $\bar{a} = \frac{1}{n}\sum_i a_i$. If we assume that all treatment-effects operate additively on each unit, i.e., through the term a_i, then we can combine this with multivariate shrinkage to get an overall mapping from experiment-level results to long-term aggregate impact, $\Delta\bar{x}$:

$$\Delta\bar{x} = (I - B - C)^{-1}\Big(\mu_t + \Sigma_t(\Sigma_t + \frac{1}{N}\Sigma_x)^{-1}(y - \mu_y)\Big).$$

Conclusion

We think it is not widely understood how the covariance of treatment effects and noise affect the interpretation of multi-outcome experiments.

In many contexts outcomes will tend to have positive covariance across units, but zero or negative covariance across different types of intervention, which has the following implications: (1) observing an unexpected positive side-effect in an experiment is *bad* news about the strength of the primary effect, (2) estimating "surrogate" outcomes using correlations across experiments will systematically over-state the strength of the surrogacy relationship, (3) estimating causal effects using correlations across experiments will have an additional bias, with ambiguous sign, (4) composite metrics will tend to have a lower signal-noise ratio than their components, and so be less-often statistically significant.

We hope in future work to clarify how broadly these results extend to other distributions, given the observation that treatment-effects are typically fat-tailed.[6]

Appendix: Derivations

We start with the most general formula for updating one vector of variables, t, having observed the realization of some other vector, y, given they have a Gaussian joint distribution:

$$V\begin{pmatrix} t \\ y \end{pmatrix} = \begin{pmatrix} \Sigma_t & \Sigma_{t,y} \\ \Sigma_{t,y}^T & \Sigma_y \end{pmatrix}.$$

Using the Schur complement we have:

$$\begin{aligned} \mathbb{E}[t|y] &= \mu_t + \Sigma_{t,y}\Sigma_y^{-1}(y - \mu_y) \\ \mathbb{V}[t|y] &= \Sigma_t - \Sigma_{t,y}\Sigma_{t,y}^{-1}\Sigma_{t,y}^T. \end{aligned}$$

Intuitively, updating to infer $\mathbb{E}[t|y]$, can be thought of in three steps: (1) we take the unexpected part of the results, $y - \mu_y$, (2) we normalize it by dividing it by its own covariance matrix, Σ_y, and (3) we transpose it into the t-space by multiplying it by the covariance between signal and truth $\Sigma_{t,y}$.[7]

When we know that y represents the results of an experiment with sample-size N, we can write:

$$\begin{aligned} y &= t + \sum_i^N x_i \\ \mathbb{V}\begin{pmatrix} t \\ y \end{pmatrix} &= \begin{pmatrix} \Sigma_t & \Sigma_t \\ \Sigma_t^T & \Sigma_t + \frac{1}{N}\Sigma_x \end{pmatrix} \end{aligned}$$

Then the optimal Bayesian inference about the treament effects, t, from the observed outcomes y, will be:

$$\mathbb{E}[t|y] = \mu_t + \Sigma_t(\Sigma_t + \frac{1}{N}\Sigma_x)^{-1}(y - \mu_y).$$

We can also write the solution with respect to the *precision* matrix, $\Phi = V\binom{t}{y}^{-1}$:

$$\mathbb{E}[t|y] = \mu_t - \sum_{j=1}^{n} \frac{\Phi_{t1,y1}}{\Phi_{t1,t1}}(y_j - \mu_{y,j}),$$

If $\Phi_{t1,y1} = 0$ then t_1 and y_1 are conditionally independent, and so y_1 has no informational content relevant to t_1.

References

Athey, Susan, Raj Chetty, Guido Imbens, and Hyunseung Kang. 2016. "Estimating Treatment Effects Using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index." *arXiv Preprint arXiv:1603.09326*.

[6] See Azevedo et al. (2019) and references therein.

[7] Note that the matrix $\Sigma_{t,y}\Sigma_y^{-1}$ represents the expected regression coefficients one would find from regressing t on y.

Azevedo, Eduardo, Alex Deng, Jose Montiel Olea, and Glen Weyl. 2019. "Empirical Bayes Estimation of Treatment Effects with Many a/b Tests: An Overview." *American Economic Review P&P*, 43–47.

Coey, Dominic, and Tom Cunningham. 2019. "Improving Treatment Effect Estimators Through Experiment Splitting." In *The World Wide Web Conference*, 285–95. ACM.

Peysakhovich, Alexander, 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. International World Wide Web Conferences Steering Committee.