Beyond Correlation: A Bayesian Multilevel Model for Effective Proxy Metric Use in A/B Tests*

Miha Gazvoda
miha.gazvoda@booking.com

Christina Katsimerou
christina.katsimerou@booking.com

October 19, 2024

1 Introduction

Designing and analyzing high-quality online experiments or A/B tests requires choosing a good primary metric that accurately reflects business results. Such metrics may be noisy or observed with significant delay. To address this issue, experimenters often use a more precisely measured proxy metric (P). Prior work [Cunningham and Kim, 2020] has demonstrated that Ps and goal (business) metrics (Gs) tend to be positively correlated across experimental units, which leads to correlated sampling errors in observed treatment effects (TEs), even in experiments without true TEs.[1] However, the correlation of true TEs may be near zero or even negative. Given the low signal-to-noise ratio (SNR) of the G, which is typical in online experiments [Deng et al., 2013], the correlation in sampling errors can mask the true correlation of TEs when examining the observed TEs of two metrics. Naively assuming that the correlation between observed TEs is equivalent to that of true TEs can lead to poor P selection. In fact, even a positive correlation between true TEs does not guarantee directional alignment between P and G [Chen et al., 2007]. Ignoring the above can result in high Type S errors [Gelman and Carlin, 2014] and negative cumulative true TEs (CTE) on the G for released (i.e. fully rolled out) experiments.

[Cunningham and Kim, 2020] used a hierarchical (multilevel) model in closed-form to adjust for correlation in sampling error and recover the joint distribution of true TEs. [Tripuraneni et al., 2024] estimated the correlation between true TEs with a probabilistic generative model to construct a composite metric from a weighted combination of Ps. Our work emphasizes that directly using the posterior distribution of the true TEs inferred from the model improves the precision of the estimated TEs and decreases the Type S error, while implicitly recovering the true correlation between TEs. Furthermore, we extend the bivariate normal model of [Cunningham and Kim, 2020] and [Tripuraneni et al., 2024] to account for heavier-than-normal tails in the population distribution of true TEs, which are frequently observed in real-world experiments [Azevedo et al., 2020].

We present results from realistic simulations to evaluate the performance of our Bayesian framework compared to standard baselines. Specifically, our approach achieves lower mean squared error (MSE) and mean absolute error (MAE) by using the posterior expectation of TEs rather than relying on observed TEs. Additionally, we compare several experiment release criteria: significance of the P, significance of the G, and two Bayesian methods based on posterior distributions of TEs for the G - one where the TE is positive with at least 95% probability, and another based on the posterior expectation being positive. Our analysis focuses on Type S error and CTE of released experiments, demonstrating the advantages of the Bayesian approach in reducing errors and improving decision-making accuracy. Our work focuses on the case of an observed but low SNR G metric, but could be easily extended to the case of unobserved G.

*Presented at the 2024 Conference on Digital Experimentation @ MIT (CODE@MIT).

[1] If P and G are positively correlated on the unit-level, then A/B test variants that happen to have more units with high P will also have more units with high G.

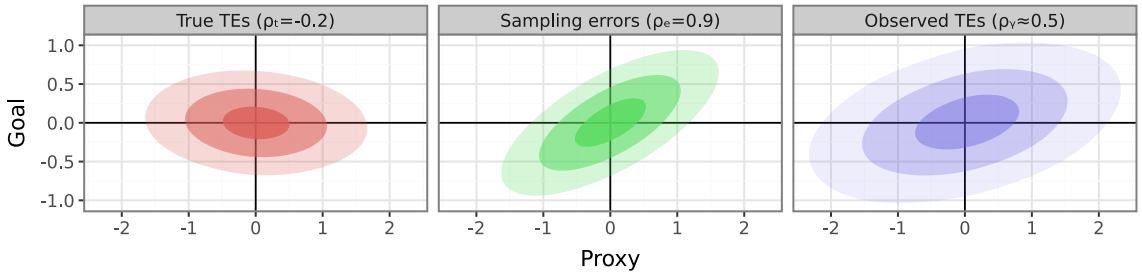

Figure 1: The panel visualizes densities of true TEs, sampling errors (note that each sampling error comes from experiment-specific covariance matrix), and observed TEs for all simulated experiments.

2 Methods

Consider a corpus of K experiments where experiment i (with $i \in \{1, ..., K\}$) has a sample size n_i. We track two metrics for each experiment: the P and G, denoted by subscripts P and G, respectively. i-th experiment has true TEs t_{P_i} and t_{G_i}, which are drawn i.i.d. from a common joint distribution with covariance matrix Σ_t. The sample sizes n_i are large enough that the sampling errors e_{P_i} and e_{G_i} follow a bivariate normal distribution, centered around the true TEs due to the Central Limit Theorem. The covariance matrix of the sampling errors Σ_{e_i} is estimated from the standard errors of observed TEs and unit-level correlation between the two metrics. The observed TEs are denoted by y_{P_i} and y_{G_i}.

We estimate the joint distribution of the true TEs using a meta-analytic approach. We assume access to a pool of homogeneous experiments with available unbiased estimates of the P and G TEs. True TEs are i.i.d. and drawn from a multivariate Student's t-distribution with ν degrees of freedom (df). With that information, we can construct a hierarchical model, resembling the one used in [Tripuraneni et al., 2024]:

$$\begin{aligned} \begin{pmatrix} t_{P_i} \\ t_{G_i} \end{pmatrix} &\sim \text{MVStudent-t}_\nu \left(\begin{pmatrix} \mu_P \\ \mu_G \end{pmatrix}, \Sigma_t \right) \\ \begin{pmatrix} y_{P_i} \\ y_{G_i} \end{pmatrix} &\sim \text{MVNormal} \left(\begin{pmatrix} t_{P_i} \\ t_{G_i} \end{pmatrix}, \Sigma_{e_i} \right) \end{aligned} \tag{1}$$

For the purposes of inference we use the plug-in estimate of Σ_{e_i}, a common practice in hierarchical modeling [Gelman et al., 1995]. Since closed-form inference in this model is not possible, we implement the model in the open-source probabilistic programming language PyMC [Abril-Pla et al., 2023]. We need to set priors for the parameters of the population distribution of TEs: df ν, locations μ_P and μ_G, scales σ_{t_P} and σ_{t_G}, and correlation ρ_t. The model simultaneously estimates the posterior distributions of the aforementioned parameters and the latent true TEs t_{P_i} and t_{G_i} of individual experiments. The posterior distribution of t_{G_i} is our primary focus, as it represents our most informed probabilistic inference about true TEs, given our assumed generative model. To make experiment release decisions, we can leverage different summary statistics from the posterior distribution of t_{G_i}. Specifically, we consider the posterior expectation and 95th percentile. We demonstrate how these statistics can inform experiment release decisions in the Results section, comparing their effectiveness to traditional significance-based approaches.

3 Results

We simulate 500 corpora, each containing 20 experiments based on the generative model described in the Methods section. This corpus size is realistic for real-world applications, assuming homogeneity across experiments. We fit the model to each corpus separately using weakly informative priors (see 3). Decreasing the corpus size affects the precision of posterior estimates, as we have less data for estimating the true parameters of the generative model.

For the P, we set both $\sigma_{e_P} = \sigma_{t_P} = 1$. For the G, we set $\sigma_{e_G} = 0.5$ and $\sigma_{t_G} = 0.25$, which could reflect a setting of upper (high SNR) funnel and goal (low SNR) metrics. The Multivariate Student's

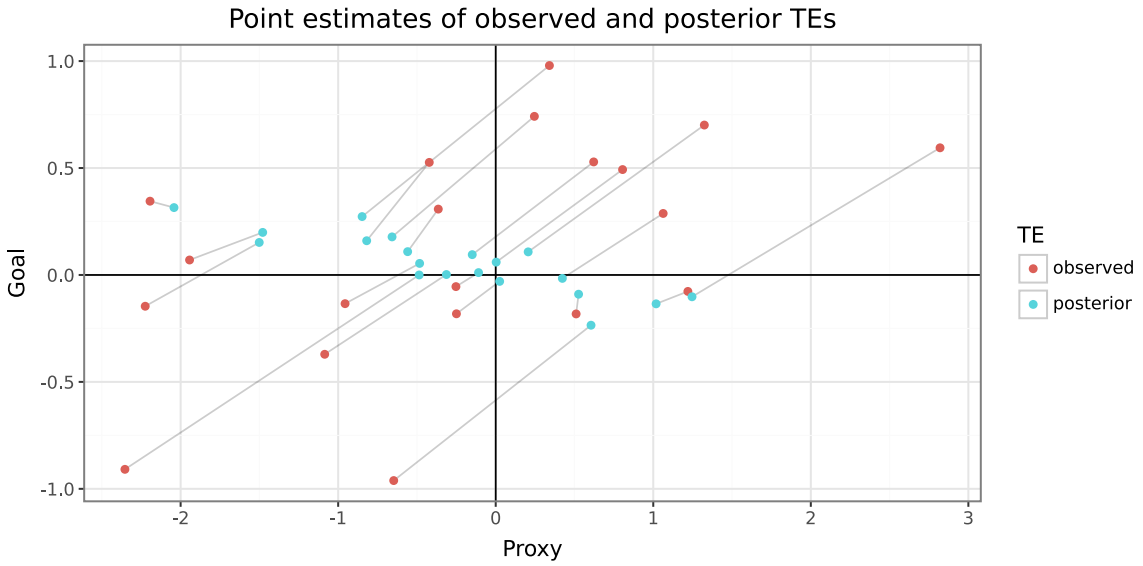

Figure 2: Comparison of observed and posterior TEs for experiments in a single corpus. The observed TEs (red) show a positive correlation ($\rho_y = 0.51$), while the posterior TEs (blue) exhibit a negative correlation, demonstrating the model's capability to infer the actual correlation of true TEs. Estimates for the same experiments are connected by lines to illustrate the effect of two dimensional shrinkage. Notably, the shrinkage doesn't consistently pull towards the centers of the true TE distribution ($\mu_P = \mu_G = 0$, denoted by dashed lines) as it would if the true TEs were not correlated or if the model lacked the ability to capture such correlations.

t-distribution of true TEs is centered at zero ($\mu_P = \mu_G = 0$)[2] with $\nu = 10$ df. We simulate different (individual) experiment sample sizes by multiplying σ_{e_P} and σ_{e_G} by a factor $k_i \sim \text{LogNormal}(0, 0.2)$. We set the correlation of the sampling errors to $\rho_e = 0.9$ and correlation of TEs to $\rho_t = -0.2$. While exaggerated to illustrate a point, these correlations remain plausible [Cunningham and Kim, 2020]. The correlation of observed TEs across all simulated experiments is $\rho_y = 0.47$. Figure 1 illustrates the densities of the simulated true TEs, sampling errors, and observed TEs of the simulated experiments.

Figure 2 compares observed and posterior TEs for individual experiments in a single corpus. The posterior estimates differ from the observed ones, even with a small corpus size. The shrinkage substantially improves the precision of TE estimates, as reflected by the large reductions in MSE and MAE for the G across all experiments, as shown in Table 1. It is important to note that these precision improvements are amplified due to the simulation setup, where true TEs are drawn from a Student t-distribution and centered at zero, matching our assumption. We expect the posterior estimates to still perform substantially better than using observed TEs alone, even when these assumptions are not entirely met in practice.

G	MSE	MAE
Observed TE	0.27	0.41
Posterior TE	0.06	0.18

Table 1: Comparison of MSE and MAE between observed and posterior TEs for G across all experiments. Improvements are also substantial for the P.

Figure 3 illustrates that the model successfully updated priors for the scale parameters and identified a negative correlation in true TEs. This finding aligns with [McElreath, 2018], which suggests that scale parameters (and especially location parameters) are generally easier to estimate than correlations. Notably, the posterior distribution for df closely matches the prior. Estimating df is often impractical due to insufficient extreme observations in smaller datasets. When extreme observations exist, using

[2]We set them directly to 0 without attempting to estimate them in our model. This is a reasonable assumption given that TEs tend to be centered close to 0.

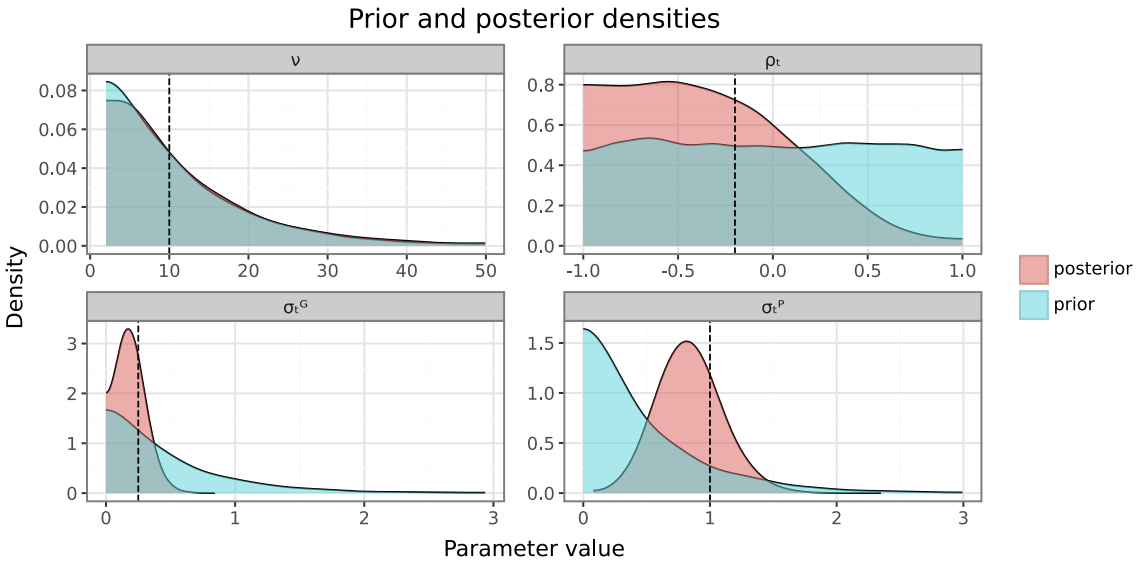

Figure 3: Prior (blue) and posterior (red) distributions of parameters with the true parameters value indicated by the dashed line. Scale and df priors are weakly informative exponential distributions.[4] The correlation ρ_t prior is flat. Posterior distributions are approximately aligned with the actual parameter values denoted by dashed vertical lines. Scales (σ_{t_P}, σ_{t_G}) are easier to estimate than correlation ρ_t. The posterior distribution for df ν closely matches the prior, which is common in smaller datasets due to insufficient number of extreme observations allowing for precise estimation.

a small df value is common to reduce their influence, as suggested by [McElreath, 2018].

We also compared 4 different release criteria: (1) significance[5] of a P as a proxy for the G, (2) significance of the G, (3) 95% posterior probability that the G's TE is positive, (4) positive posterior expectation for the G. We fitted our model to multiple experiments in parallel. However, in real-world scenarios, decisions are made sequentially. Therefore, we can envision release criteria based on posterior probabilities in our simulation as if we were making decisions for each experiment individually, using information from the other 19 experiments in our corpus as if they were past experiments.

As shown in Table 2, using the P as a primary metric leads to a higher release rate (13%) but results in a negative CTE (-80) and a high Type S error (60%). Directly targeting the G decreases the release rate to 8% but improves the CTE (95) and reduces the Type S error to 17%. Due to shrinkage, the 95% positive posterior criterion achieves the lowest release rate (2%) but successfully controls for Type S error (5%). Using positive expected posterior TEs as a release criterion yields the highest CTE (680) but comes with a higher risk of Type S error (29%). These results demonstrate that posterior-based methods allow experimenters to trade-off between risk (Type S error) and potential gains (CTE), providing more flexibility in decision-making compared to traditional significance-based approaches.

Criteria	Release Rate	CTE	Type S Error
Sig. P	13%	-80	60%
Sig. G	8%	220	17%
Post. 95% G	2%	95	5%
Post. G	50%	680	29%

Table 2: Comparison of release criteria in terms of release rate (the proportion of experiments with a release decision), CTE on the G, and Type S error.

[5]We used a two-sided significance level of 0.1 for the false positive rate.

4 Conclusions

Evaluating a P based solely on correlation of observed TEs in experiments often overstates the strength of the relationship between the P and G. Our simulations demonstrate how this can lead to poor decision-making, resulting in the negative CTE of released experiments and a high Type S error. Employing an appropriate probabilistic model addresses these issues and reduces MAE and MSE compared to observed TEs. Additionally, we extended the model from prior work to account for fat-tailed TEs, which are common in real-world experiments. The same model can be easily extended to predict the posterior distribution of TEs for the G when it is not observed by treating it as missing data. While our simulations are based on a generative model, we believe this approach is valuable in real-world scenarios where assumptions may not perfectly hold. To validate its robustness, we recommend conducting sensitivity analyses to explore potential model misspecifications and their implications.

References

[Abril-Pla et al., 2023] Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesbeck, C. J., Kochurov, M., Kumar, R., Lao, J., Luhmann, C. C., Martin, O. A., et al. (2023). Pymc: a modern, and comprehensive probabilistic programming framework in python. *PeerJ Computer Science*, 9:e1516.

[Azevedo et al., 2020] Azevedo, E. M., Deng, A., Montiel Olea, J. L., Rao, J., and Weyl, E. G. (2020). A/b testing with fat tails. *Journal of Political Economy*, 128(12):4614–000.

[Chen et al., 2007] Chen, H., Geng, Z., and Jia, J. (2007). Criteria for surrogate end points. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(5):919–932.

[Cunningham and Kim, 2020] Cunningham, T. and Kim, J. (2020). Interpreting experiments with multiple outcomes.

[Deng et al., 2013] Deng, A., Xu, Y., Kohavi, R., and Walker, T. (2013). 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*, pages 123–132.

[Gelman and Carlin, 2014] Gelman, A. and Carlin, J. (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on psychological science*, 9(6):641–651.

[Gelman et al., 1995] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

[McElreath, 2018] McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

[Tripuraneni et al., 2024] Tripuraneni, N., Richardson, L., D'Amour, A., Soriano, J., and Yadlowsky, S. (2024). Choosing a proxy metric from past experiments. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5803–5812.