Breaking the Winner's Curse with Bayesian Hybrid Shrinkage

Richard Mudd, Rina Friedberg, Ilya Gorbachev, Houssam Nassif, and Abbas Zaidi

Meta Inc
{rmudd, rinafriedberg ilyagorbachev, houssamn, abbaszaidi}@meta.com

1 Introduction

A "Winner's Curse" arises in large-scale online experimentation platforms [1] when the same experiments are used to both select treatments and evaluate their effects. In these settings, classical difference-in-means estimators of treatment effects are upwardly biased and conventional confidence intervals are rendered invalid [2]. The bias scales with the magnitude of sampling variability and the selection threshold. It decreases inversely with the treatment's true effect size [3].

Bayesian methods that leverage empirical priors have been proposed and demonstrated, in specific settings, to yield superior inferential properties under selection [4, 5] when compared to the standard "Face Value" estimators. However, Bayesian analysis is sensitive - and arguably especially susceptible under selection [6] - to the choice of prior, and models that require computationally expensive numerical integration techniques are ill-suited for at-scale deployment [7].

We propose a new Bayesian approach that incorporates experiment-specific "local shrinkage" factors that mitigate sensitivity to the choice of prior and improve robustness to assumption violations. Crucially, we demonstrate how the associated posterior distribution can be estimated without numerical integration techniques, making it a practical choice for at-scale deployment.

2 Modeling Selection in Bayesian Analyses

In a frequentist analysis, the Winner's Curse is an artifact of the sampling distribution of a statistic being altered by selection. This implies that an explicit correction for selection via a selection model is always required [2].

The need for a selection model under the Bayesian paradigm, is less obvious. One longstanding perspective is that selection is not required because posterior distributions are already conditioned on the data [8]: If a different sample of data was observed under a hypothetical replication, this is irrelevant with respect to the posterior at hand. This conditioning does not necessarily account for the selection mechanism that led to those observations though.

The interaction of the selection mechanism with the parameter space determines whether such an adjustment is needed in a Bayesian analysis [9]. If we consider parameters and data to be sampled from a joint distribution, no explicit selection adjustment is required. On the other hand, if we consider a parameter to be sampled from its marginal distribution and held fixed, with the data then sampled repeatedly from the associated conditional distribution, an adjustment is required. This abstraction is best illustrated by example [10]:

- Suppose we consider a set of experiments that test different changes to a product. In each experiment, the data collected is used to make a launch decision according to some predetermined criteria. For experiments that launched, we want to estimate the effect size. In this context a Bayesian approach would *not* require adjustment for selection.
- Suppose we consider a single proposed change to a recommender system, that we decide to test multiple times in different experiments. If we are interested in using the set of experiments that pass some launch criteria to estimate the size of the proposed change, we *would* need to adjust for selection.

Different settings leading to different solutions arise from the likelihood principle: likelihoods that differ only by a scalar multiplier are equivalent, and should lead to same inferences.

Theorem 2.1. *With joint selection, the posterior distribution under selective inference is equivalent to its unadjusted counterpart, eliminating the need for a separate selection model.*

Proof. Under joint selection, let the joint distribution of $\{\hat{\theta}, \theta\}$ be given by $\pi_S(\hat{\theta}, \theta) = \pi(\theta) \cdot f(\hat{\theta}|\theta) \cdot \frac{1(\hat{\theta} \in S)}{\pi(S)}$, where the marginal selection probability is $\pi(S) = \int \int_S \pi(\hat{\theta}|\theta)\pi(\theta)\mathrm{d}\hat{\theta}\mathrm{d}\theta$. The marginal density of the data under selection is $m(\hat{\theta}) = \int f(\hat{\theta}|\theta)\pi(\theta)\mathrm{d}\theta$. Truncated to the selected sample S, this marginal density is $\int 1(\hat{\theta} \in S)\pi(\theta)\frac{f(\hat{\theta}|\theta)}{m(S)}\mathrm{d}\theta = \frac{m(\hat{\theta})}{m(S)} \cdot 1(\hat{\theta} \in S)$. Next consider the posterior under selection $\pi_S(\hat{\theta}|\theta) = \frac{\pi_S(\hat{\theta},\theta)}{m_S} = \frac{\pi(\theta)\cdot f(\hat{\theta}|\theta)}{m(\hat{\theta})} = \pi(\theta|\hat{\theta})$ which is identical to the unadjusted posterior. □

Many applied settings can reasonably be described by the joint distribution paradigm, meaning that a well-motivated prior is often sufficient to overcome the Winner's curse. In this context, strategies to validate the modeling choice and ensure a well-calibrated prior distribution are vital.

3 Formulation

3.1 A Bayesian Model for Inference Under Selection

Consider a collection of N experiments indexed by $i = 1, \ldots, N$. Each experiment consists of m units, indexed by $j = 1, \ldots, m$, which are assigned to treatment conditions denoted by Z_{ji}. $Z_{ji} = 1$ indicates that unit j in experiment i receives the treatment, while $Z_{ji} = 0$ indicates control. The outcome of interest for each unit is Y_{ji}. Our estimand for each experiment i is the ratio of the expected potential outcomes under treatment and control, expressed as $\theta_i = \frac{E[Y(1)]}{E[Y(0)]}$.

The classical or *"Face Value"* estimator, $\hat{\theta}_i^{FV}$, is computed as the ratio of sample averages of observed outcomes in treated and control groups:

$$\hat{\theta}_i^{FV} = \frac{\frac{1}{\sum_{j=1}^m Z_{ji}} \sum_{j=1}^m Y_{ji} Z_{ji}}{\frac{1}{\sum_{j=1}^m (1 - Z_{ji})} \sum_{j=1}^m Y_{ji}(1 - Z_{ji})}$$

with standard error $\hat{\sigma}_i$. This estimator is known to be biased in the presence of selection effects.

To address this, we propose a new *"Bayesian Hybrid Shrinkage"* approach, which can be formulated as a hierarchical post-hoc model:

$$\begin{aligned} \hat{\theta}_i | \theta_i, \hat{\sigma_i}^2 &\sim \mathrm{N}(\theta_i, \hat{\sigma_i^2}), \\ \theta_i | m_0, \lambda_i, \tau &\sim \mathrm{N}(m_0, \lambda_i \cdot \tau), \\ \lambda_i | a, b &\sim \text{InverseGamma}\left(\frac{a}{2}, \frac{b}{2}\right). \end{aligned}$$

True effect θ_i is assigned a normal prior with mean m_0 and variance composed of a global scale parameter τ modulated by local shrinkage factor λ_i, which allows the model to adaptively shrink estimates differently across experiments. λ_i is an inverse-Gamma hyperprior parameterized by a and b, which control the shrinkage distribution across experiments. The hierarchical structure balances borrowing strength across experiments with flexibility to accommodate experiment-specific variability.

3.2 Inference

Inference can be carried out either via posterior simulation—using iterative sampling over the target and nuisance parameters—or analytically by fixing λ_i at its posterior mode. The posterior distribution of the true effect θ_i, conditional on the observed estimator $\hat{\theta}_i$, the local shrinkage factor λ_i, and the global scale parameter τ, is given by:

$$\theta_i \mid \hat{\theta}_i, \lambda_i, \tau \sim \mathrm{N}\left(\frac{\hat{\sigma}_i^2}{\hat{\sigma}_i^2 + \lambda_i\tau} m_0 + \frac{\lambda_i\tau}{\hat{\sigma}_i^2 + \lambda_i\tau}\hat{\theta}_i, \quad \left(\frac{1}{\hat{\sigma}_i^2} + \frac{1}{\lambda_i\tau}\right)^{-1}\right). \tag{1}$$

A special case of this posterior arises when the local shrinkage factor is fixed at $\lambda_i = 1$ for all experiments, corresponding to a Bayesian estimator that imposes only global shrinkage [5]. We refer to this as the *"Bayesian Global Shrinkage"* model.

3.3 Validation via Predictive Checking

Given the complexities of validation in experimental settings, we take inspiration from *predictive checking* [11, 12]. This type of assessment uses predictive simulation of new data under a model of interest M that we denote as $\hat{\theta}_{rep}$, along with a related statistic $T(\cdot)$ to characterize discrepancies against its observed counterpart $g(T(\hat{\theta}_{rep}), T(\hat{\theta}))$. This discrepancy can be used to assess any aspect of the model we want to validate (e.g., prior parameter choices or decision boundary) facilitated by the construction of a reference distribution, $p(g(T(\hat{\theta}_{rep}), T(\hat{\theta}))|M, \hat{\theta})$.

These flexible constructions enable assessment of quantities like coverage for uncertainty intervals or goodness-of-fit of the posterior distribution. Specifically, to understand goodness-of-fit, one intuitive quantity is the tail area probability $g(T(\hat{\theta}_{rep}), T(\hat{\theta})) = T(\hat{\theta}_{rep}) \geq T(\hat{\theta})$, an analogue to a p-value that can be computed for a given model averaged over the posterior distribution of the parameter of interest $p(\theta_i|M, \hat{\theta})$. This concept can be augmented by data splitting strategies such as replication studies or data fission [13]. We utilize this concept in the following section to assess performance first in simulation and then empirically from real experiment data.

4 Results

4.1 Simulated Experiments

To evaluate the performance of the proposed *Bayesian Hybrid Shrinkage* approach, we conduct simulated experiments comparing it against the *Face Value* and *Bayesian Global Shrinkage* approaches. These simulations are designed to reflect plausible real-world scenarios where the analysis prior is misspecified in various ways:

1. **Misspecified mean:** The analysis prior mean $m_0 = 0$ differs from the true effect size distribution mean, where $\theta_i \sim \mathrm{N}(\mu, \epsilon)$.

2. **Heavy-tailed distributions:** The true effect sizes follow a t-distribution with ν degrees of freedom, $\theta_i \sim t_\nu(\mu, \epsilon)$, which has heavier tails than the assumed normal analysis prior.

3. **Hidden Selection:** True effects are drawn as two-dimensional vectors from a bivariate normal distribution, $\boldsymbol{\theta_i} \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the covariance matrix $\boldsymbol{\Sigma}$ has correlation ρ. Selection is applied jointly on both dimensions, but the analysis considers only one target parameter, testing robustness to unmodeled correlated selection.

The results of these simulations are summarized in Figure 1, which compares the three approaches across the three simulation settings and three performance metrics: *Mean Squared Error (MSE)*, *Bias*,

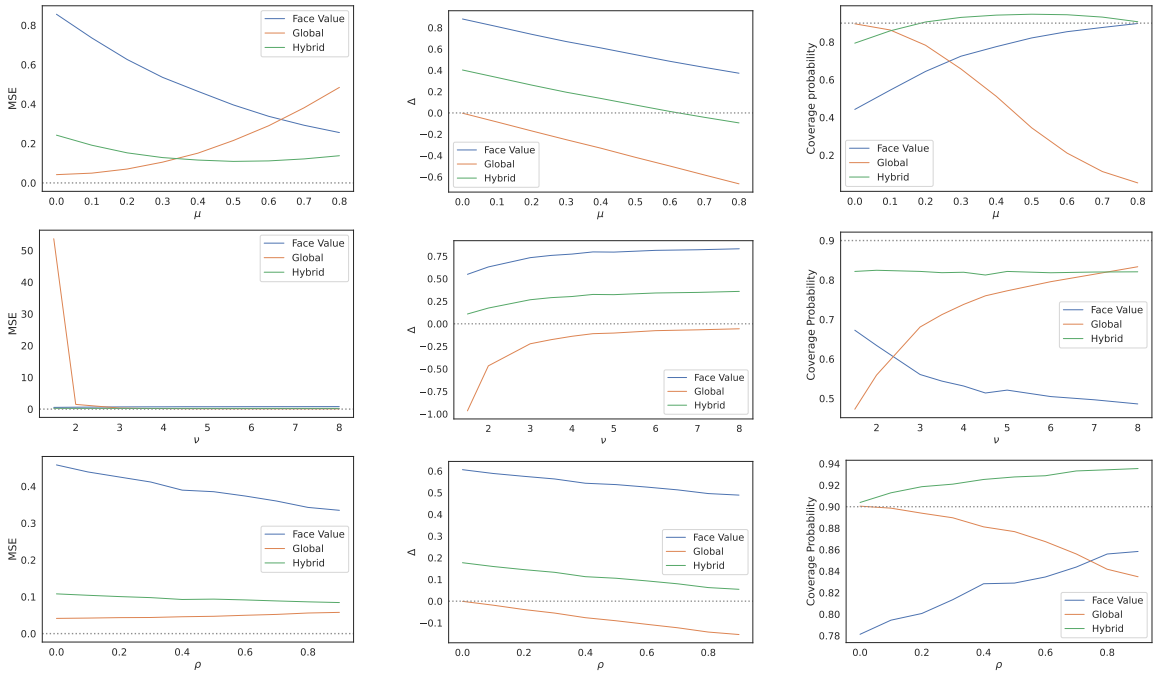

Figure 1: MSE (left), bias (Δ, center), and coverage probability (right) for the *Face Value* (blue), *Bayesian Global Shrinkage* (orange), and *Bayesian Hybrid Shrinkage* (green) approaches, as a function of prior mean (μ, top row), degrees of freedom (ν, middle row), and correlation (ρ, bottom row).

and *Coverage Probability* of 90% uncertainty intervals. As expected, the *Bayesian Global Shrinkage* approach performs optimally when the prior is correctly specified. However, the *Bayesian Hybrid Shrinkage* approach is shown to be more robust to misspecification and consistently outperforms the *Face Value* approach in all settings considered.

4.2 Empirical Analysis

We further evaluate the three approaches in a real-world setting, analyzing 167 experiments with paired replication studies. We compare the mean absolute error (MAE) and the coverage probability for 90% confidence/credible intervals. The results, summarized in Table 1, demonstrate that the *Bayesian Hybrid Shrinkage* approach outperforms both the *Face Value* and *Bayesian Global Shrinkage* methods in terms of MAE, while maintaining strong coverage properties.

Model	MAE	Coverage
Face Value	1.280	0.92
Global Shrinkage	1.121	0.91
Hybrid Shrinkage	1.012	0.91

Table 1: Summary of Performance Metrics – Coverage and (MAE x 1000) – across a collection of real-world experiments with paired replication studies.

5 Conclusion

This paper introduce a two stage Bayesian Shrinkage estimator to tackle the Winner's Curse with a scalable strategy for posterior inference. Future research plans include: expanding to additional validation strategies like prior elicitation to achieve specific operating characteristics; guidance for practical implementation of these techniques; offline-evaluation priors [14]; deeper formalization of theoretical properties for posterior inference. We anticipate that deeper exploration along these avenues will enable a wider adoption of these methods in the online experimentation domain.

References

[1] Minyong R Lee and Milan Shen. Winner's curse: Bias estimation for total effects of features in online controlled experiments. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 491–499, 2018.

[2] Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on winners. *The Quarterly Journal of Economics*, 139(1):305–358, 2024.

[3] Erik W van Zwet and Eric A Cator. The significance filter, the winner's curse and the need to shrink. *Statistica Neerlandica*, 75(4):437–452, 2021.

[4] Simon Ejdemyr, Martin Tingley, Yian Shang, and Travis Brooks. Estimating the returns from an experimentation program. In *ACIC Conference*, 2024.

[5] Ryan Kessler. Overcoming the winner's curse: Leveraging bayesian inference to improve estimates of the impact of features launched via a/b tests. 2024.

[6] Daniel García Rasines and G Alastair Young. Bayesian selective inference. In *Handbook of Statistics*, volume 47, pages 43–65. Elsevier, 2022.

[7] Zhaoqi Li, Houssam Nassif, and Alex Luedtke. Estimation of subsidiary performance metrics under optimal policies. *Statistica Sinica*, 37(3), 2027.

[8] AP Dawid. Selection paradoxes of bayesian inference. *Lecture Notes-Monograph Series*, pages 211–220, 1994.

[9] Daniel Yekutieli. Adjusted bayesian inference for selected parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(3):515–541, 2012.

[10] Spencer Woody, Oscar Hernan Madrid Padilla, and James G Scott. Optimal post-selection inference for sparse signals: a nonparametric empirical bayes approach. *Biometrika*, 109(1):1–16, 2022.

[11] Donald B Rubin. More powerful randomization-based p-values in double-blind trials with non-compliance. *Statistics in medicine*, 17(3):371–385, 1998.

[12] Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996.

[13] James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data fission: splitting a single data point. *Journal of the American Statistical Association*, pages 1–12, 2023.

[14] Mohamed A Radwan, Quinn Lanners, Jiasheng Zhang, Serkan Karakulak, Houssam Nassif, and Murat Ali Bayir. Counterfactual evaluation of ads ranking models through domain adaptation. In *Workshops of Conference on Recommender Systems (RecSys)*, 2024.