Comparing The Bayesian and Frequentist Procedures for False Discovery Control: A Simulation Study

Simulation Design

To provide a direct comparison between the Bayesian procedure in our proposed framework and standard frequentist procedures, we conducted a simulation study designed to replicate the statistical environment of anomaly replication tests. We simulated 10,000 correlated t-statistics from a mixture of null and non-null hypotheses. The true proportion of nulls was set to 0.60, so that 40% of tests corresponded to genuine signals.

Dependence was introduced through an equicorrelated structure with correlation coefficient $\rho = 0.30$, a level chosen to reflect the strong comovement often observed among characteristic-sorted portfolios. Each statistic was generated with 360 degrees of freedom, corresponding approximately to the number of months in a typical 30-year sample adjusted for factor regressors. For the non-null cases, the signal strength was governed by a noncentrality parameter of $\delta_{\text{alt}} = 1.50$, which creates a weak signal regime where the relative conservativeness of the methods has the greatest practical impact. The data-generating mechanism took the form

$$T_i = \frac{\mu_i + G_i}{\sqrt{W_i/\nu}},$$

where G captures the equicorrelated Gaussian component, , W_i is chi-squared with $\nu = 360$ degrees of freedom, and μ_i is zero under the null and equal to δ_{alt} under the alternative.

False Discovery Controlling Procedures

We compared three procedures side by side. The first benchmark was the 3-sigma $|t|$ rule, which declares a discovery whenever the absolute t-statistic exceeds 3, following the convention advocated by Harvey et al. (2016). The second benchmark was the Benjamini–Hochberg (BH) procedure at a false discovery rate (FDR) level of $\alpha = 0.05$, applied to the two-sided p-values of the simulated statistics, in line with how the replication literature evaluates multiple testing (Chordia etal., 2020; Hou et al., 2020). The third method was the Bayesian decision rule in our framework, which rejects the null when the posterior probability that the anomaly abnormal return is non-positive falls below $pr^c = 0.05$. To compute these probabilities efficiently for thousands of tests, we converted each p-value into a signed z-equivalent using the sign of the simulated t-statistic and then applied the closed-form likelihood expression derived in Section 3.1 of our manuscript. This yields the posterior probability of null hypothesis given a prior null probability π_{null} and a prior scale parameter τ. We fixed $\tau = 1.00$ and examined three prior skepticism levels, $\pi_{\text{null}} \in \{0.30, 0.50, 0.70\}$, to illustrate how the Bayesian rule flexibly adapts to different beliefs about the prevalence of true effects.

FDR Estimation

For the frequentist procedures, we estimated the realized false discovery rate using Storey's plug-in estimator (Story, 2002, 2003). Specifically, we estimated the proportion of true null hypotheses as the proportion of p-values above a tuning parameter λ divided by $1 - \lambda$, averaged across $\lambda = 0.5$, 0.8, and 0.9, and then plugged this estimate into the expression $\widehat{\text{FDR}} = \hat{\pi}_{\text{null}} m \tau^* / R$, where τ^* is the rejection threshold implied by the procedure and R is the number of rejections. For the Bayesian method, the FDR is defined directly as the mean posterior probability of null hypothesis across all discoveries, which equals the Bayes FDR

derived in Equation (7) of our manuscript. This measure is immediately interpretable as the expected proportion of false positives among declared anomalies.

Result Reporting and Discussion

Table 1 presents the results.

Table 1: Head-to-head comparison between The Bayesian and frequentist procedures

π_{null}	Procedure	Tests	Discoveries	$\widehat{\text{FDR}}$	Overlap–B
0.30	Bayesian	10,000	**3,046**	**0.0224**	–
–	3-sigma $\lvert t\rvert$	10,000	554	0.0301	554
–	BH	10,000	512	0.0307	512
0.50	Bayesian	10,000	**1,597**	**0.0248**	–
–	3-sigma $\lvert t\rvert$	10,000	554	0.0301	554
–	BH	10,000	512	0.0307	512
0.70	Bayesian	10,000	**704**	**0.0272**	–
–	3-sigma $\lvert t\rvert$	10,000	554	0.0301	554
–	BH	10,000	512	0.0307	512

As shown in Table 1, the Bayesian procedure yielded more discoveries than both BH and the 3-sigma rule across all scenarios, while maintaining a lower or comparable false discovery rate. At a prior skepticism of $\pi_{\text{null}} = 0.50$, the Bayesian rule identified 1,597 discoveries with a Bayes FDR of 2.48%. In contrast, the 3-sigma and BH procedures identified only 554 and 512 discoveries, with estimated FDRs of 3.01% and 3.07%, respectively. When the prior skepticism was reduced to $\pi_{\text{null}} = 0.30$, the Bayesian rule identified 3,046 discoveries with a Bayes FDR of 2.24%. At the more skeptical setting $\pi_{\text{null}} = 0.70$, the Bayesian rule identified 704 discoveries with a Bayes FDR of 2.72%. In all cases, the set of rejections made by the

frequentist methods was a strict subset of those made by the Bayesian rule. These results demonstrate that the Bayesian procedure is consistently less conservative than the frequentist benchmarks in a dependent, heavy-tailed environment, while still maintaining control of false discoveries.

It is important to note that the apparent "best" performance of the Bayesian decision rule in Table 1 at $\pi_{\text{null}} = 0.30$, where it yields the highest number of discoveries at the lowest Bayes FDR, ***does not*** imply that 0.30 is the "best" prior skepticism setting. The specification of π_{null} in empirical applications reflects the researcher's prior beliefs about the plausibility of anomalies *before examining the data*, not a parameter to optimize based on observed outcomes. Different researchers may choose different prior values informed by theory or domain expertise. Our goal here is to demonstrate how the Bayesian procedure behaves across a plausible range of priors. The main conclusion is that, regardless of the prior setting, it is less conservative and consistently delivers more discoveries than the frequentist benchmarks while maintaining lower or comparable false discovery rates.

Code and reproducibility

We include the complete R script that generates the table above: `simulate_procedure_comparison.R` in this repository.

Reference

Storey, J.D., 2002. A direct approach to false discovery rates. J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (3), 479–498.

Storey, J.D., 2003. The positive false discovery rate: A Bayesian interpretation and the q-value. Ann. Stat. 31 (6), 2013–2035.