Tweedie's Formula for Testing: Score Identities from Hypothesis Tests to Diffusion Models

Anonymous Authors[1]

Abstract

Tweedie's formula expresses the posterior mean under Gaussian noise as the observation shifted by the score of the log-marginal density. The same identity connects two developments that have evolved separately: Rubin's decision-theoretic view of Bayes hypothesis testing and the denoising step at the heart of modern diffusion generative models. We develop the testing side. Under a spike-and-slab hypothesis, the Tweedie posterior mean decomposes into posterior signal probability times conditional signal mean. The associated local false discovery rate rule is equivalent to Bayes-factor thresholding and satisfies an oracle screening consistency theorem. The key observation is that in a diffusion model with Gaussian forward process, the score network's denoised prediction equals the posterior signal mean under the training prior. For out-of-distribution detection, the training score can be used to recover the training log density, which can then be compared with a specified null density; the fdr-BF equivalence supplies principled false discovery rate thresholds without heuristic calibration.

1. Introduction

One of the deepest tensions in modern statistics is the discrepancy between frequentist p-values and Bayesian measures of evidence. Berger & Sellke (1987) showed that, under equal prior odds, a p-value of 0.05 implies a posterior probability of the null of at least 0.29 for symmetric unimodal alternatives (more than five times the nominal significance level).

A single identity connects these perspectives. For $z \mid \theta \sim N(\theta, 1)$ with prior $g(\theta)$ and marginal $m(z) = \int \phi(z - \theta) g(\theta)\, d\theta$, differentiating the marginal gives

$$m'(z) = m(z)\big(\mathbb{E}[\theta \mid z] - z\big), \tag{1}$$

or equivalently, with $\ell(z) = \log m(z)$, *Tweedie's formula* (Robbins, 1956; Efron, 2011):

$$\mathbb{E}[\theta \mid z] = z + \ell'(z). \tag{2}$$

Efron (2011) exploits (2) for estimation after selection: the bias of a selected estimator is corrected by $\ell'(z)$. This paper develops the testing side: the same identity yields the posterior signal statistic $T(z) = z + \ell'(z)$, while the associated local-fdr/Bayes-factor rule gives the Bayes decision threshold and connects to classical thresholding rules.

Independently, diffusion generative models (Ho et al., 2020; Song et al., 2021) train a score network $s_\theta(x_t, t) \approx \nabla \log p_t(x_t)$ to reverse a Gaussian noise process. The denoising prediction is

$$\hat{x}_0 = \frac{x_t + \sigma_t^2\, s_\theta(x_t, t)}{\alpha_t}, \tag{3}$$

which is again the Tweedie formula applied to $x_t \mid x_0 \sim N(\alpha_t x_0, \sigma_t^2)$. Thus $T(x_t) = x_t + \sigma_t^2 s_\theta(x_t, t) = \alpha_t \hat{x}_0$ is the posterior signal mean $\alpha_t\, \mathbb{E}[x_0 \mid x_t, H_1]$: the score network computes a signal-size statistic, and the training log density needed for OOD Bayes factors can be recovered from the score by path integration when normalization is calibrated.

This paper makes three contributions. *(C1)* We derive the Tweedie posterior signal statistic for spike-and-slab testing and show it equals the product of the posterior signal probability and the conditional signal mean (Proposition 1). *(C2)* We prove the fdr-Bayes factor equivalence and an oracle screening consistency theorem (Theorem 3). *(C3)* We show that a score network computes the Tweedie posterior signal mean $T(x_t) = \alpha_t\, \mathbb{E}[x_0 \mid x_t, H_1]$ (Proposition 4), and that comparing the score-derived training density with a specified null density provides principled fdr thresholds for OOD detection via Corollary 2.

2. Related Work

Robbins 1956 introduced compound decision theory; Efron 2011; 2012 uses the log-marginal score $\ell'(z)$ for bias correction after selection and large-scale estimation. Our

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

paper develops the testing side: $T(z) = z + \ell'(z)$ is a posterior signal-size statistic coupled to the Bayes-optimal local-fdr/Bayes-factor rule, rather than only an estimator of θ. Datta and Polson 2025 characterize log-marginals for which the Tweedie posterior mean has a polynomial form.

The p-value/Bayes-factor discrepancy is established by Berger & Sellke (1987); Corollary 2 makes it precise. The BH procedure (Benjamini & Hochberg, 1995) and empirical Bayes thresholding (Johnstone & Silverman, 2004) are the classical benchmarks; Theorem 3 connects the fdr threshold to the Donoho-Johnstone 1994 universal threshold via the Bayes factor K_n.

Score matching (Hyvärinen, 2005; Vincent, 2011) underlies diffusion models (Ho et al., 2020; Song et al., 2021). Graham et al. 2023 use reconstruction error $\|x_t - \alpha_t \hat{x}_0\|^2$ with heuristic thresholds; this discards x_t and uses only score magnitude, whereas $T(x_t)$ combines both; Corollary 2 supplies the principled threshold.

3. The Tweedie Test Statistic

We work with the canonical Gaussian sequence model: $Z_i \mid \theta_i \sim N(\theta_i, 1)$, independently for $i = 1, \ldots, n$. Under the spike-and-slab prior $g(\theta) = \pi_0 \delta_0(\theta) + \pi_1 g_1(\theta)$, the marginal is $m(z) = \pi_0 \phi(z) + \pi_1 m_1(z)$, where $m_1(z) = \int \phi(z-\theta) g_1(\theta)\, d\theta$ and $\ell_0(z) = \log \phi(z)$, $\ell(z) = \log m(z)$.

For $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$, the Bayes decision is determined by the posterior odds, equivalently by the local fdr or Bayes factor. Rubin's 1971 quadratic-plus-flat loss assigns cost θ^2 for accepting H_0 when the true signal is θ, and a flat cost c for rejection regardless of θ. The Bayes-optimal rule rejects when $\mathbb{E}[\theta^2 \mid z] > c$; since $\mathbb{E}[\theta^2 \mid z] = T(z)^2 + \mathrm{Var}(\theta \mid z)$ by the law of total variance, thresholding $|T(z)|$ approximates the Bayes-optimal decision. The score identity (1) gives the following explicit form.

Proposition 1 (Tweedie posterior signal statistic)**.** *Under the spike-and-slab prior, the marginal score decomposes as*

$$\ell'(z) = \Pr(H_0 \mid z) \cdot (-z) + \Pr(H_1 \mid z) \cdot \ell_1'(z),$$

and the Tweedie posterior signal statistic satisfies

$$T(z) \;=\; \ell'(z) + z \;=\; \Pr(H_1 \mid z) \cdot \mathbb{E}[\theta \mid z, H_1]. \quad (4)$$

The local false discovery rate and Bayes factor are

$$\mathrm{fdr}(z) = \pi_0\, e^{\ell_0(z) - \ell(z)}, \qquad B_{10}(z) = e^{\ell_1(z) - \ell_0(z)}. \quad (5)$$

Equation (4) reads: $T(z)$ vanishes when the posterior probability of the alternative is small or when the expected signal is zero, and grows with both the evidence for H_1 and the magnitude of the signal. The parallel with Efron's formula is exact: estimation uses $\mathbb{E}[\theta|z] - z = \ell'(z)$ to shrink toward the prior; testing uses $T(z) = \ell'(z) + z$ to measure posterior signal strength.

Corollary 2 (fdr-BF equivalence)**.** *The rule* $\mathrm{fdr}(z) \le q$ *is equivalent to* $B_{10}(z) \ge K$*, where* $K = (\pi_0/\pi_1)(1-q)/q$*. At* $q = 0.05$ *and* $\pi_0 = 0.9$*, the required evidence is* $B_{10}(z) \ge 171$*, far above the Bayes factor implied by a* p*-value of* 0.05*.*

This corollary makes the Berger-Sellke bound precise: a p-value of 0.05 corresponds to a Bayes factor near 1 for many priors, while fdr control at 5% requires overwhelming evidence.

Fix $q \in (0, 1)$ and define $\widehat{S} = \{i : \mathrm{fdr}(Z_i) \le q\}$. Under mild conditions on the slab g_1 (symmetry, monotone Bayes factor, and score decay $\ell_1'(z) = o(1)$ as $|z| \to \infty$), the following theorem shows that this procedure finds all strong signals and makes no false discoveries.

Theorem 3 (Oracle screening consistency)**.** *Let* $K_n = (\pi_0/\pi_1)(1-q)/q$*,* $t_n = \sqrt{2 \log K_n}$*, and* $s_n = |S_0|$*. Suppose* $\log K_n - (1+\varepsilon)\log n \to +\infty$ *for some* $\varepsilon > 0$*.*

(i) *No false discoveries:* $\Pr(\widehat{S} \subseteq S_0) \to 1$.

(ii) *No missed strong signals: if* $\min_{i \in S_0} |\theta_i| \ge t_n + c_n$ *with* $s_n \Phi(-c_n) \to 0$*, then* $\Pr(S_0 \subseteq \widehat{S}) \to 1$.

(iii) *Asymptotic normality: for each detected signal,* $(Z_i - \theta_i) \Rightarrow N(0, 1)$.

The threshold $t_n = \sqrt{2 \log K_n}$ is the Tweedie analogue of the Donoho-Johnstone universal threshold $\sqrt{2 \log n}$ (Donoho & Johnstone, 1994): both arise from balancing Gaussian tail decay against a multiplicity factor. When $\pi_1 = c/n$, t_n recovers the Donoho-Johnstone rate; the theorem's rate condition ($\log K_n \gg (1+\varepsilon)\log n$) requires π_1 decaying faster than $n^{-(1+\varepsilon)}$.

4. Score Networks and the Tweedie Posterior Mean

We now show that the Tweedie posterior signal statistic is the quantity diffusion models compute during denoising, and that this connection provides a decision-theoretic foundation for score-based OOD detection.

A diffusion model with Gaussian forward process (Ho et al., 2020; Song et al., 2021) defines a family of marginals indexed by noise level $t \in [0, 1]$:

$$x_t = \alpha_t x_0 + \sigma_t \varepsilon, \quad \varepsilon \sim N(0, I),$$

so $x_t \mid x_0 \sim N(\alpha_t x_0,\, \sigma_t^2 I)$ and the marginal is $p_t(x_t) = \int N(x_t;\, \alpha_t x_0,\, \sigma_t^2 I)\, p_0(x_0)\, dx_0$.

Score matching (Hyvärinen, 2005; Vincent, 2011) trains $s_\theta(x_t, t) \approx \nabla_{x_t} \log p_t(x_t)$ by minimizing $\mathbb{E}\big[\|s_\theta(x_t,t) - \nabla_{x_t} \log p_t(x_t \mid x_0)\|^2\big]$, which is equivalent to the denoising objective $\mathbb{E}\big[\|s_\theta(x_t,t) + \varepsilon/\sigma_t\|^2\big]$.

Proposition 4 (Score network computes the Tweedie posterior mean)**.** *Under the Gaussian forward process, Tweedie's formula applied to $x_t \mid x_0 \sim N(\alpha_t x_0, \sigma_t^2 I)$ gives*

$$\mathbb{E}[x_0 \mid x_t] = \frac{x_t + \sigma_t^2 \nabla_{x_t} \log p_t(x_t)}{\alpha_t}. \tag{6}$$

A score network trained on p_0 approximates $\nabla_{x_t} \log p_t^{train}(x_t)$, so its denoised prediction $\hat{x}_0 = (x_t + \sigma_t^2 s_\theta)/\alpha_t$ satisfies

$$T(x_t) \;=\; x_t + \sigma_t^2\, s_\theta(x_t, t) \;=\; \alpha_t\, \mathbb{E}[x_0 \mid x_t,\, H_1]. \tag{7}$$

Thus $T(x_t)$ is the posterior signal mean under H_1; it is large when the training posterior assigns large signal magnitude to the observed x_t. The Bayes-optimal decision for H_0 vs H_1 uses $\log B_{10}(x_t) = \log p_t^{train}(x_t) - \log p_t^{null}(x_t)$, where $\log p_t^{train}$ can be recovered from s_θ by path integration when the density normalization is calibrated (Song et al., 2021).

Proof. For $x_t \mid x_0 \sim N(\alpha_t x_0, \sigma_t^2 I)$, the master identity (1) applied with noise variance σ_t^2 gives $\nabla_{x_t} \log p_t(x_t) = (\mathbb{E}[\alpha_t x_0 \mid x_t] - x_t)/\sigma_t^2$, which rearranges to (6). Equation (7) follows: $s_\theta \approx \nabla \log p_t^{\text{train}}$, so $T = x_t + \sigma_t^2 s_\theta \approx \alpha_t \mathbb{E}[x_0 \mid x_t, H_1]$. ■ □

For OOD detection, define H_0: test point from p_{null}, H_1: from training distribution p_0. The log Bayes factor is $\log B_{10}(x_t) = \log p_t^{\text{train}}(x_t) - \log p_t^{\text{null}}(x_t)$, where $\log p_t^{\text{train}}$ can be recovered from the score network by path integration when the density normalization is calibrated (Song et al., 2021), and p_t^{null} is specified (e.g. Gaussian null). By Corollary 2, thresholding $B_{10}(x_t) \geq K$ targets posterior local fdr level q across the test batch, and Theorem 3 gives screening consistency in the sparse Gaussian sequence setting when signals are strong enough.

Remark 1. *Existing diffusion-based OOD detectors (Graham et al., 2023) use reconstruction error $\|x_t - \alpha_t \hat{x}_0\|^2 = \sigma_t^4 \|s_\theta\|^2$, which discards x_t and uses only the score magnitude. $T(x_t) = x_t + \sigma_t^2 s_\theta$ combines both, while the Bayes factor obtained from the training and null densities via Corollary 2 provides the principled threshold without cross-validation.*

5. Experiments

We validate the score-based OOD detection framework of Section 4 and briefly summarize the sparse-testing results.

We consider the 1-D bimodal mixture $p_0 = 0.5\, N(-3, 1) + 0.5\, N(3, 1)$ at noise level $\alpha_t = 0.88$, $\sigma_t = 0.475$. The OOD distribution is $p_{\text{null}} = N(0, 0.3^2)$ (near-null, no structured signal). Three methods detect in-distribution points at nominal level $q = 0.10$, evaluated on 1000 in-distribution and 1000 OOD test points. The principled BF threshold is derived analytically from Corollary 2 ($K = 9$ at equal prior odds); heuristic thresholds are calibrated to the 10th percentile of each statistic on a held-out in-distribution validation set.

Table 1. OOD detection comparison ($q = 0.10$, $K = 9$). FDP: false discovery proportion (fraction of in-distribution calls that are OOD). Heuristic thresholds require labeled calibration data; the principled BF threshold K follows from Corollary 2.

Method	AUC	FDP	Power	Threshold
Principled BF (Cor. 2)	0.985	0.005	0.859	$B_{10} \geq K$
Rec. err. $\|x_t - \alpha_t \hat{x}_0\|^2$	0.770	0.383	0.912	Heuristic
Obs. norm $\|x_t\|$	0.985	0.027	0.926	Heuristic

The principled BF method achieves AUC $= 0.985$ and FDP $= 0.005$, deriving its threshold $K = 9$ from Corollary 2 without any calibration data. Reconstruction error ($\|x_t - \alpha_t \hat{x}_0\|^2 = \sigma_t^4 \|s_\theta\|^2$), the score-norm proxy used by existing diffusion-based OOD detectors (Graham et al., 2023), achieves lower AUC (0.770) and severely inflated FDP (0.383): the score s_θ vanishes at the symmetric null center by design, so OOD points near $x_t = 0$ have near-zero reconstruction error and are misclassified as in-distribution. The Tweedie statistic $T(x_t) = x_t + \sigma_t^2 s_\theta$, by contrast, uses both the score correction and the raw observation, remaining large for structured in-distribution signals in this symmetric mixture. The observation norm $\|x_t\|$ achieves the same AUC in this toy example but provides no theoretical FDP guarantee; in harder problems (in-distribution close to null, or high-dimensional x_t) the heuristic threshold can fail, whereas Corollary 2 gives an analytic posterior-fdr target.

A three-layer MLP trained by denoising score matching on the same p_0 achieves T-statistic AUC $= 0.984$; using the analytic BF threshold gives FDP $= 0.001$, matching the analytical oracle (Figure 1). In the classical sparse-signal setting ($n = 2000$, $\pi_1 = 0.10$, signals from $N(3, 1)$), Tweedie/fdr achieves low FDP *per replication* (FDP $= 0.01$, SD $= 0.01$) by targeting the posterior local fdr, while BH, which controls FDR only in expectation, shows higher FDP variability (FDP $= 0.09$, SD $= 0.03$, regularly exceeding $q = 0.10$ in individual runs); the power difference (0.37 vs. 0.57) reflects the conservativeness of the local-fdr threshold.

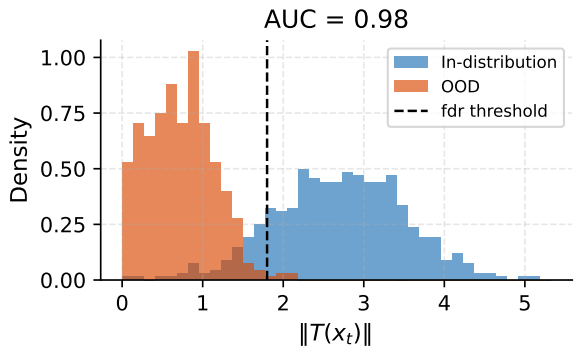

Figure 1. Distribution of $\|T(x_t)\|$ for in-distribution (blue) and OOD (orange) points. The dashed line is a visual reference: the smallest displayed $\|T(x_t)\|$ among points selected by the BF rule $B_{10} \geq K$. A three-layer MLP trained by score matching achieves AUC $= 0.984$, matching the analytical oracle at AUC $= 0.98$.

6. Discussion

The Tweedie formula $T(z) = z + \ell'(z)$ sits at the intersection of classical decision theory and modern generative AI. As a posterior signal statistic coupled to a local-fdr/Bayes-factor rule, it connects the Bayes factor, the local fdr, the Donoho-Johnstone threshold, and the Benjamini-Hochberg procedure. As the denoised prediction of a score network, it encodes the posterior signal mean; the training density needed for OOD Bayes factors can be recovered from the score, and Corollary 2 supplies the principled fdr threshold without heuristic calibration.

Three extensions are immediate. First, the framework extends to composite nulls via the generalized score decomposition, covering one-sided tests and nuisance parameters. Second, e-values (Polson et al., 2026) can replace p-values in the sequential setting, where the anytime-validity of score-martingale tests connects to the Tweedie framework through the log-marginal excess $\ell(z) - \ell_0(z)$. Third, the Gaussian assumption can be relaxed: Tweedie's formula extends to any natural exponential family, and the score matching objective extends to non-Gaussian noise schedules in diffusion models.

References

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300, 1995.

Berger, J. O. and Sellke, T. Testing a point null hypothesis: The irreconcilability of P values and evidence. *Journal of the American Statistical Association*, 82(397):112–122, 1987.

Datta, J. and Polson, N. G. Polynomial log-marginals and Tweedie's formula: When is Bayes possible? *arXiv:2509.05823*, 2025.

Donoho, D. L. and Johnstone, I. M. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

Efron, B. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Efron, B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2012.

Graham, M. S., Pinaya, W. H. L., Tudosiu, P.-D., Nachev, P., Ourselin, S., and Cardoso, M. J. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Johnstone, I. M. and Silverman, B. W. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, 32(4):1594–1649, 2004.

Polson, N. G., Sokolov, V., and Zantedeschi, D. Bayes, E-values and testing. *arXiv:2602.04146*, 2026.

Robbins, H. An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1:157–163, 1956.

Rubin, H. A decision-theoretic approach to the problem of testing a null hypothesis. In Le Cam, L., Neyman, J., and Scott, E. L. (eds.), *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 359–372. University of California Press, Berkeley, 1971.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.