A DENSITY-RATIO READOUT THEORY OF LABEL-FREE OUT-OF-DISTRIBUTION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) detection is not simply a test of nominal atypicality. Reference typicality need not reveal which distribution a sample was drawn from. An unlabeled test batch adds population-level evidence of how the test distribution differs from the reference. This paper asks what becomes detectable from this information and how much can be recovered from finite data. At the population level, the batch-to-reference density ratio identifies where the test population is overrepresented relative to the reference, providing direct evidence of distributional difference. Finite data, however, limit how finely this ratio can be resolved, and finer resolution requires higher-order batch moments. Low-degree readouts that maximize the deflection, a standardized batch–reference contrast, have closed forms in the batch moments under a whitened Gaussian working model, with degree 2 requiring only the batch mean and covariance. Experiments from CIFAR to ImageNet show that the simple degree-2 readout outperforms ratio estimators on average, with its near-OOD gain largely reproduced by the directions along which the batch covariance departs from the reference.

1 INTRODUCTION

What should an out-of-distribution (OOD) detector detect? A test point can be out-of-distribution in two different senses. It can be atypical under the in-distribution (ID) law, or it can provide evidence that it was drawn from a different distribution altogether. These notions need not agree. If a region contains more test points than the reference predicts, points in that region are likely drawn from a different distribution, even when the region is common under the reference.

A detector built from ID data alone scores atypicality under the reference and has no view of whether a region is overrepresented under a different law. Overrepresentation compares the probability of a region under the test population with its probability under the reference. Such a detector can estimate the probability under the reference but not under the test population. Its score for a query therefore depends on where the query lies under the reference, not on how many test points share its region. If anomalies fall in a high-density region of the reference, such a detector may treat them as ID.

Much of modern deep OOD detection operates in this reference-only regime. Beginning with confidence scores from ID-trained classifiers (Hendrycks & Gimpel, 2017), subsequent work has constructed scores from classifier logits (Liu et al., 2020), feature-space distances (Lee et al., 2018; Sun et al., 2022), and generative likelihoods (see Lu et al., 2025, for a recent survey). While their mechanisms differ, they typically score each query independently using a fixed ID-trained model, optionally augmented with representations or statistics constructed from ID data. This regime is fundamentally limited: reference-distribution atypicality alone does not in general characterize OOD (Farquhar & Gal, 2022), while OOD detection is not learnable from ID data alone without assumptions on the OOD distribution (Fang et al., 2022).

Access to the test batch opens a different information regime. The unlabeled test batch can be compared with the ID reference to detect OOD samples (Garg et al., 2023). The batch distribution then provides information about the samples being scored, so the detector can use evidence that a sample was drawn from a different distribution, beyond its atypicality under the reference. Even anomalies in a high-density region of the reference can then be flagged, so long as that region is overrepresented in the batch. This overrepresentation is measured pointwise by the batch-to-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

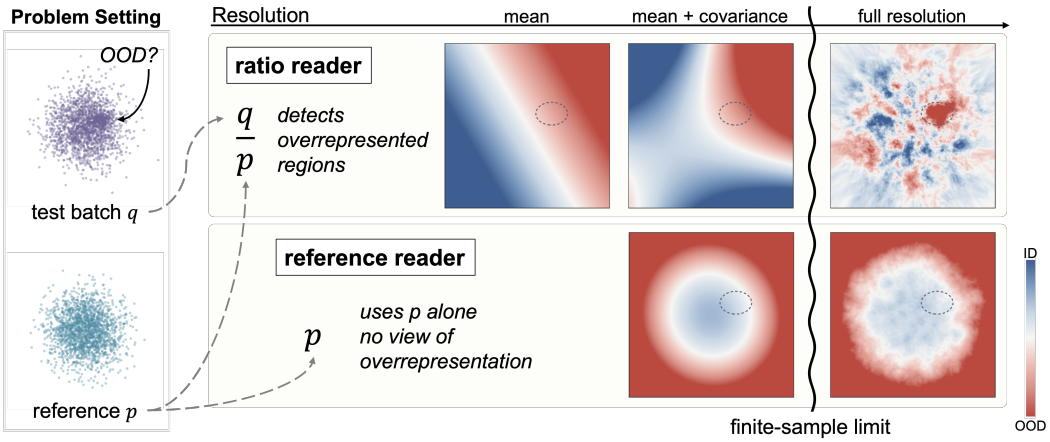


Figure 1: Comparing the test batch with the reference reveals anomalies that the reference alone can miss. The reference p contains ID samples, whereas the unlabeled test batch q may also contain hidden anomalies. Here, the anomalies lie in a high-density region of p (dashed ellipses), so a reference reader may treat them as ID. A ratio reader instead detects that the same region is overrepresented in q relative to p. Rows distinguish the estimand, q/p versus p, and columns increase the resolution. Moving right increases expressivity but also the finite-sample cost of estimation.

reference density ratio, which exceeds one where the batch holds more points than the reference predicts.

Yet this ratio must be estimated from finite data, which may not support estimating it at full resolution. For OOD detection, it is not enough to detect that the test distribution differs from the reference. The detector must recover enough of the ratio to rank individual samples.

We therefore organize label-free OOD readouts along two axes (Figure 1): *what a detector reads* (the *estimand*) and *how finely it reads it* (the *resolution*). A reference reader scores the reference density, whereas a ratio reader scores the batch-to-reference density ratio. Resolution is indexed by polynomial degree, which sets the highest order of batch moments a readout uses. Higher degrees approximate the ratio more finely but require more samples. At each degree, we use the readout that maximizes the deflection, a standardized batch–reference contrast. Within this framework, we compare the two estimands at the population level, derive closed-form low-degree readouts under a whitened Gaussian working model, and examine the finite-sample cost of each degree. On CIFAR and ImageNet benchmarks, we evaluate these readouts and decompose the degree-2 readout to locate its OOD signal.

Our contributions are:

- **A two-axis view of OOD readouts.** We organize label-free OOD detection by *estimand* and *resolution*: whether a detector reads the reference density or a batch-to-reference density ratio, and how finely that ratio can be resolved from a finite unlabeled batch. This view explains what each estimand can and cannot reveal, and separates detecting a batch shift from recovering its sample ranking.
- **Closed-form finite-resolution readouts.** Under a whitened Gaussian working model, explicit moment-based scores are derived. The degree-2 readout g_2 is a closed form in the batch mean and covariance with no tuning parameters. On the evaluated benchmarks, this simple readout already matches or exceeds ratio estimators, with its gain concentrated on near-OOD.
- **The near-OOD signal lies in covariance directions.** This near-OOD gain arises from the directions along which the batch covariance departs from the reference. Scoring these directions alone recovers this gain, which persists after the mean direction is projected out.

2 RELATED WORK

Batch-based OOD scoring and density-ratio estimation. Test-time OOD methods use unlabeled observations in different ways: Fan et al. (2024) adapt an existing score through linear regression, Zhang et al. (2026) augment a feature memory bank with pseudo-labeled test samples, and OODD maintains a dynamic dictionary of test features (Yang et al., 2025). The nominal-plus-contaminated sampling structure also underlies semi-supervised novelty detection, where the problem reduces to two-sample classification (Scott & Blanchard, 2009; Blanchard et al., 2010). OOD detection with wild data instead uses an auxiliary unlabeled mixture of ID and OOD samples at training time (Katz-Samuels et al., 2022; Du et al., 2024). Direct density-ratio estimation fits the population ratio without separately estimating the two densities (Kanamori et al., 2009; Sugiyama et al., 2012), including for per-sample outlier scoring (Hido et al., 2011). DDE learns a critic from a dual KL objective and uses its pointwise outputs for OOD scoring (Garg et al., 2023). POT scores the unlabeled batch by a geometric transport cost to ID class prototypes (Ke et al., 2025). We instead maximize the deflection over nested polynomial spaces. This makes resolution explicit in both the score's dependence on the query and the batch moments required to construct it, allowing us to separate the contributions of successive moment orders.

Restricted polynomial readouts. Restricting the readout to successive polynomial degrees is a classical idea in detection theory and statistics. In signal processing, a deflection-optimal Volterra detector expresses the optimal finite-order statistic by projecting the displaced likelihood ratio onto the Volterra subspace (Picinbono & Duvaut, 1990). Related constructions appear in restricted least-squares ratio fitting (Kanamori et al., 2009), smooth goodness-of-fit tests (Neyman, 1937; Ledwina, 1994), and low-degree likelihood-ratio methods (Hopkins, 2018; Kunisky et al., 2022). We instead use restricted polynomial readouts to rank individual samples and examine the finite-sample cost of estimating each order.

3 A DENSITY-RATIO READOUT THEORY

Reference density alone need not yield an optimal OOD ranking, whereas the population density ratio does. We characterize finite-resolution ratio readouts and their estimation costs.

3.1 SETTING

Let $z = \phi(x) \in \mathbb{R}^d$ be a feature representation, where ϕ may be a pretrained encoder or the identity map. A *reference sample* is drawn i.i.d. from a nominal density p of ID examples, and an unlabeled *test batch* has density

$$q = (1 - \pi)\, p + \pi\, a, \qquad \pi \in (0, 1), \tag{1}$$

where a is the unknown anomalous (OOD) density and π the contamination rate. A detector may use the query, the reference, and the entire test batch; larger scores indicate stronger evidence of an anomaly.

3.2 ESTIMANDS

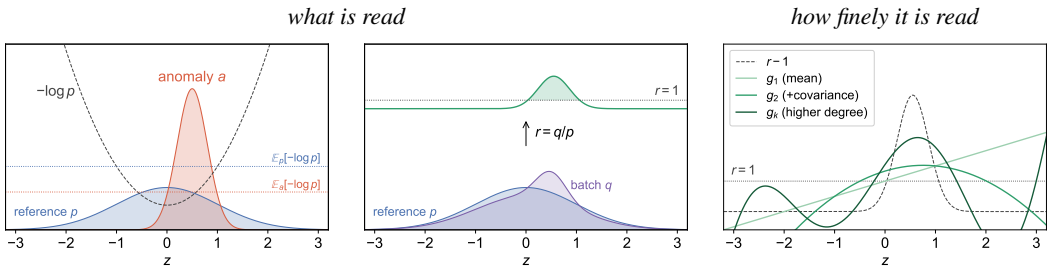
3.2.1 POPULATION TARGET

The density ratio $q(z)/p(z)$ measures where the batch is unexpectedly dense relative to the reference, and it is the classical optimum. Writing

$$r = \frac{q}{p} = (1 - \pi) + \pi \frac{a}{p}, \qquad \Pr(\text{ood} \mid z) = \frac{\pi a}{q} = 1 - \frac{1 - \pi}{r}, \tag{2}$$

r is a positive affine transform of a/p and the posterior is strictly increasing in r, so q/p, a/p, and the posterior OOD probability induce the same ordering on the support of p. By the Neyman–Pearson lemma (Neyman & Pearson, 1933), thresholding any of the three attains the population-optimal ROC curve (Scott & Blanchard, 2009; Blanchard et al., 2010).

We distinguish *reference readers*, which target p, from *ratio readers*, which target q/p.



(a) reference reader $s_p = -\log p$ (b) ratio reader $r = q/p$ (c) resolution $g_k = \mathrm{Proj}_{V_k}(r-1)$

Figure 2: A one-dimensional illustration. (a) *Estimand*: a concentrated anomaly in a high-density region of the reference receives a lower $-\log p$ score on average (dotted lines: mean scores $\mathbb{E}_p[s_p]$ and $\mathbb{E}_a[s_p]$), so the mean gap reverses, $G_p < 0$ (Proposition 1). (b) The density ratio $r = q/p$ instead reveals the same region as overrepresented in the batch and cannot invert at population level (Section 3.2.1). (c) *Resolution*: $g_k = \mathrm{Proj}_{V_k}(r-1)$ is the optimal degree-$\leq k$ readout for deflection (Theorem 1); g_1 captures mean shift and g_2 adds covariance shift. Higher k approximates more of r, but the sample cost grows with the degree.

3.2.2 Entropy competition

Many detectors, including the unlabeled Mahalanobis distance (Rippel et al., 2021), one-sample kNN (Sun et al., 2022), and class-conditional Gaussian-mixture likelihoods (Morteza & Li, 2022), estimate p rather than q/p. If anomalies cluster in a high-density region of p, they score as more typical than the reference. The population separation of a score is the anomaly–reference mean gap

$$G(s) := \mathbb{E}_a[s] - \mathbb{E}_p[s]. \tag{3}$$

A positive gap means higher average anomaly scores, not AUROC above chance.

Because each estimand determines a ranking only up to strictly monotone transformations, we focus on two canonical scores with closed-form mean gaps: $s_p = -\log p$ and $s_r = \log(a/p)$. Any strictly decreasing function of p produces the same ranking as s_p, and any strictly increasing function of a/p (equivalently q/p) produces the same ranking as s_r. Other scores that target the same estimand inherit that qualitative behavior, but need not obey the identities below (Appendix B.1).

Proposition 1 (entropy and divergence decomposition)**.** *Write $\Delta H := H(p) - H(a)$, assuming the displayed entropies and cross-entropies are finite. Let $G_p := G(s_p)$ and $G_r := G(s_r)$. Then*

$$G_p = \mathrm{KL}(a\|p) - \Delta H \qquad \textit{for the canonical reference reader } s_p = -\log p, \tag{4}$$

$$G_r = J(a,p) \geq 0 \qquad \textit{for the canonical ratio reader } s_r = \log(a/p). \tag{5}$$

Proof. (i) Follows by linearity of expectation and the definitions of H and KL; (ii) because $G_r = \mathrm{KL}(a\|p) + \mathrm{KL}(p\|a)$, the decomposition defining Jeffreys' divergence (Jeffreys, 1946; Kullback & Leibler, 1951). □

Equation (4) restates the expected-likelihood decomposition of Caterini & Loaiza-Ganem (2021) as a mean gap of $s_p = -\log p$. It gives attenuation when $0 < \Delta H < \mathrm{KL}(a\|p)$, zero mean gap at equality, and a negative gap when $\Delta H > \mathrm{KL}(a\|p)$. The entropy term cancels in the ratio (Figure 2a,b).

3.3 Resolution

The population ratio is the target, and the resolution is the part of that target that a restricted family can express. We first study population readouts, then their finite-sample cost.

3.3.1 Optimal polynomial readouts

Consider the nested spaces $V_k = \{f : \deg f \leq k,\ \mathbb{E}_p f = 0\} \subset L^2(p)$ of centered polynomials. Assume p has finite moments through degree $2k$, q through degree k, and the inner product $\langle f_1, f_2\rangle_p := \mathbb{E}_p[f_1 f_2]$ is nondegenerate on V_k. Write $\|f\|_p := \sqrt{\langle f, f\rangle_p}$.

The population ratio is optimal for ranking, but over the restricted class V_k, direct AUROC maximization becomes a nonlinear pairwise-ranking problem. We therefore seek a tractable finite-resolution criterion that preserves the $L^2(p)$ geometry of the ratio. The observable mean contrast between the batch and the reference is

$$\Delta(s) := \mathbb{E}_q[s] - \mathbb{E}_p[s] = \pi G(s). \tag{6}$$

If F_p is the continuous reference c.d.f. of a score s, then

$$\Delta(F_p \circ s) = \pi\big\{\mathrm{AUROC}(a, p; s) - \tfrac{1}{2}\big\}. \tag{7}$$

Thus, up to the contamination factor π, AUROC excess over chance is exactly a batch–reference mean contrast after reference-rank calibration (Appendix B.1). We therefore retain this two-sample contrast but optimize it on the raw score: the rank calibration is score-dependent and generally leaves V_k, whereas the raw contrast preserves its linear $L^2(p)$ geometry. Normalizing by the reference standard deviation removes arbitrary score scale, giving the deflection (Picinbono, 1995),

$$D(f) := \frac{\Delta(f)}{\mathrm{sd}_p(f)}. \tag{8}$$

This criterion is observable from the unlabeled batch and reference and turns finite-resolution optimization into an $L^2(p)$ projection of the excess density ratio. If $q \ll p$ and $r = q/p \in L^2(p)$, then $\mathbb{E}_q[f] = \mathbb{E}_p[rf]$, so $\Delta(f) = \langle f, r-1\rangle_p$ for $f \in V_k$. The optimal direction is therefore the part of the excess ratio that this class can express.

Theorem 1 (deflection optimality)**.** *Let $L(f) = \Delta(f)$. There is a unique $g_k \in V_k$ such that $L(f) = \langle f, g_k\rangle_p$ for every $f \in V_k$.*

(i) *If $g_k \neq 0$, then for every $f \in V_k$ with $\mathrm{sd}_p(f) > 0$,*

$$D(f) \;\le\; \|g_k\|_p, \tag{9}$$

with equality if and only if $f = cg_k$ for some $c > 0$. If $g_k = 0$, then $D(f) = 0$ for every such f.

(ii) *If $q \ll p$ and $r = q/p \in L^2(p)$,*

$$g_k = \mathrm{Proj}_{V_k}(r - 1). \tag{10}$$

The proof of Theorem 1 is in Appendix B.

Remark 1 (Classical restricted readouts). On the same class and $L^2(p)$ geometry, g_k is the population object of deflection-optimal Volterra detection (Picinbono & Duvaut, 1990), low-degree likelihood-ratio projection (Hopkins, 2018; Kunisky et al., 2022), unregularized least-squares density-ratio fitting (Kanamori et al., 2009), and the MMD witness whose kernel is the reproducing kernel of V_k under $\langle\cdot,\cdot\rangle_p$ (Gretton et al., 2012).

3.3.2 Closed forms

After reference whitening (Appendix A.3), the Gaussian working model $p = \mathcal{N}(0, I_d)$ decomposes V_k into orthogonal Hermite shells (Appendix B.3). The first captures the mean shift; the second adds the second-moment shift, which splits into radial and directional parts. The first two projections of Theorem 1 therefore have closed forms (Figure 2c). Write $m := \mathbb{E}_q[z]$ and $S := \mathbb{E}_q[zz^\top] - I$. Then

$$g_1(z) = m^\top z, \qquad g_2(z) = m^\top z + \tfrac{1}{2}\langle S,\; zz^\top - I\rangle_{\mathrm{F}}, \tag{11}$$

with $\|g_1\|_p^2 = \|m\|^2$ and $\|g_2\|_p^2 = \|m\|^2 + \frac{1}{2}\|S\|_F^2$, where $\langle\cdot,\cdot\rangle_{\mathrm{F}}$ denotes the Frobenius inner product.

The factor $\frac{1}{2}$ is the Gaussian second-moment Gram weight. The contrast $\Delta(g_2)$, however, does not need Gaussianity: for any whitened reference,

$$\mathbb{E}_q[g_2] - \mathbb{E}_p[g_2] \;=\; \|m\|^2 + \tfrac{1}{2}\|S\|_F^2 \;\ge\; 0, \tag{12}$$

with equality if and only if $m = 0$ and $S = 0$; this follows by substituting the definitions of m and S, since whitening gives $\mathbb{E}_p[g_2] = 0$. The general Gram form is in Appendix B.3.

Remark 2 (Degree-2 equivalence). Under the Gaussian working model, the common readout of Remark 1 is (11), up to an additive or positive multiplicative constant. An RBF-kernel MMD witness (Gretton et al., 2012) lives in a richer class and recovers a different representer.

What the quadratic readout captures. Degree 1 detects only mean shifts. At degree 2 the excess second moment $S = \frac{\operatorname{tr} S}{d} I + S_0$ splits into trace and traceless parts,

$$g_2(z) = \underbrace{m^\top z}_{\text{mean}} + \underbrace{\tfrac{\operatorname{tr} S}{2d}\big(\|z\|^2 - d\big)}_{\text{radial}} + \underbrace{\tfrac{1}{2}\langle S_0, zz^\top\rangle_{\mathrm{F}}}_{\text{directional}}, \tag{13}$$

so the readout can respond to mean-preserving second-moment changes: the radial term is the whitened Mahalanobis score up to a batch-dependent sign, and the directional term captures anisotropic shifts in $S = \mathbb{E}_q[zz^\top] - I$.

3.3.3 Moment estimation

For an unlabeled test batch $\{z_j\}_{j=1}^n \subset \mathbb{R}^d$, write $\widehat{M}_\ell = n^{-1}\sum_{j=1}^n z_j^{\otimes \ell}$ for its empirical moments: $\widehat{M}_1$ is the batch mean and $\widehat{M}_2$ the second-moment matrix.

Definition 1 (degree-$\leq k$ readout)**.** *A detector T has degree $\leq k$ if (i) its batch dependence factors through $(\widehat{M}_1, \ldots, \widehat{M}_k)$, and (ii) with those moments fixed, $z \mapsto T(z)$ is a polynomial of degree $\leq k$.*

The empirical plug-in of g_k is such a readout. For g_1, g_2, replace m, S in Equation (11) by moments of the other $n-1$ batch points. Each leave-one-out score satisfies Definition 1 with these $n-1$ moments held fixed. This leave-one-out construction removes query self-alignment (Appendix C), but not sampling error. These moments update per arrival, so a stream can be treated as a growing batch (Appendix C).

An unstructured degree-k shell has $\Theta(d^k)$ coefficients, and an order-of-magnitude calculation suggests that detecting a shift in it requires $n \asymp d^{k/2}$ batch samples (Appendix B.5; see Yan & Zhang, 2023 for related rigorous asymptotics). Higher degree therefore trades approximation error for estimation cost; no fixed degree is uniformly preferable at finite sample size. Neighborhood ratios lie outside every fixed-degree class (Appendix B.4).

4 Experiments

4.1 Protocol

We use the OpenOOD v1.5 split (Yang et al., 2022; Zhang et al., 2023) in the unlabeled batch setting of Section 3: each test batch mixes ID-test and OOD, and a detector may use the whole batch. The ID datasets are CIFAR-10, CIFAR-100, and ImageNet-1K. For CIFAR, the main grid keeps one near-OOD and one far-OOD set; results on the full OpenOOD v1.5 grid are in Appendix D.3.

The primary encoders are supervised ViT-B/16 (Dosovitskiy et al., 2021) ($d = 768$) and self-supervised DINOv2 ViT-S/14 (Oquab et al., 2024) ($d = 384$). We report AUROC in the main text and FPR@95 in Appendix D.3. Each reported value is a 5-seed mean of one (encoder, reference, OOD) pair at a stated (n, π), where each seed draws a reference bank of $n_{\mathrm{ref}} = 5{,}000$ samples and one test batch. For the main comparison, we use $n = 2{,}000$ and $\pi = 0.1$. Experimental details are in Appendix D.1.

The compared readers are g_1, g_2, $\widehat{r}_{\mathrm{kNN}}$, and DDE (Garg et al., 2023), which read q/p, and Mahalanobis, Mahalanobis++ (Müller & Hein, 2025), and one-sample kNN, which read p. "Label-free" refers to the detection stage; the encoder may be supervised. All readers are label-free in this sense except Mahalanobis++, which uses ID class labels and serves as a labeled reference reader. Mahalanobis is the unlabeled single-Gaussian score, and DDE trains a critic on the test batch and the reference. Detector definitions and external baselines are in Appendix D.2.

4.2 What the test batch adds

Table 1 compares reference readers with ratio readers, all computed from the same features and reference bank. On average, g_2 is the strongest reader, and its margin is concentrated on near-OOD. There it stays ahead of the class-conditional Mahalanobis++ and of the learned ratio estimator DDE. RMDS (Ren et al., 2021), the strongest reference-only baseline, stays below g_2 on average (0.901 vs. 0.949 on the full grid; Table 7).

Table 1: AUROC on the OpenOOD main grid ($n = 2{,}000$, $\pi = 0.1$, 5-seed means). g_1 and g_2 are closed-form leave-one-out polynomial readouts; $\widehat{r}_{\mathrm{kNN}}$ is the leave-one-out local ratio. Red marks inversion; bold marks the best value in each row. Further baselines and FPR@95 are in Appendix D.3.

				reference density p			density ratio q/p			
ref.	encoder	OOD	type	Maha.	Maha++	kNN	g_1	g_2	$\widehat{r}_{\mathrm{kNN}}$	DDE
CIFAR-10	ViT-B/16	SVHN	far	0.895	0.879	0.918	**1.000**	**1.000**	**1.000**	**1.000**
		CIFAR-100	near	0.859	0.878	0.892	0.781	0.880	**0.893**	0.873
	DINOv2	SVHN	far	0.881	0.955	0.917	**1.000**	**1.000**	**1.000**	**1.000**
		CIFAR-100	near	0.813	0.853	0.852	0.784	0.855	**0.884**	0.797
CIFAR-100	ViT-B/16	SVHN	far	0.432	0.423	0.635	0.999	**1.000**	**1.000**	**1.000**
		CIFAR-10	near	0.747	0.790	0.713	0.808	**0.902**	0.859	0.843
	DINOv2	SVHN	far	0.513	0.760	0.661	**1.000**	**1.000**	**1.000**	**1.000**
		CIFAR-10	near	0.654	0.735	0.652	0.776	0.843	**0.848**	0.733
ImageNet-1K	ViT-B/16	iNaturalist	far	0.918	0.949	0.830	0.994	**0.996**	0.991	0.992
		Textures	far	0.859	0.882	0.847	0.945	**0.956**	0.901	0.947
		OpenImage-O	far	0.893	0.907	0.822	0.871	**0.928**	0.777	0.918
		SSB-hard	near	0.644	0.694	0.559	0.671	**0.728**	0.521	0.686
		NINCO	near	0.827	0.869	0.737	0.834	**0.912**	0.614	0.881
	DINOv2	iNaturalist	far	0.980	0.987	0.969	0.994	**0.996**	0.994	0.989
		Textures	far	0.886	0.900	0.868	0.956	**0.976**	0.953	0.940
		OpenImage-O	far	0.833	**0.918**	0.900	0.780	0.851	0.850	0.782
		SSB-hard	near	0.712	0.678	0.633	0.571	**0.716**	0.678	0.622
		NINCO	near	0.694	0.797	0.756	0.719	**0.857**	0.740	0.789
Mean, far-OOD (10)				0.809	0.856	0.837	0.954	**0.970**	0.947	0.957
Mean, near-OOD (8)				0.744	0.787	0.724	0.743	**0.836**	0.755	0.778
Mean, all (18)				0.780	0.825	0.787	0.860	**0.911**	0.861	0.877

The largest gap between the two estimands appears on CIFAR-100/SVHN (ViT-B/16). There Mahalanobis scores 0.432 and Mahalanobis++ scores 0.423, while g_2 reaches 1.000. An AUROC below 0.5 indicates that SVHN features lie closer to the reference center than ID test features do. Consistent with Proposition 1, a reference reader ranks such points as typical, whereas a ratio reader identifies them as anomalous because they are overrepresented in the batch relative to the reference.

Additional baselines, FPR@95, and results on the full OpenOOD v1.5 grid are in Appendix D.3, and sensitivity to the kNN neighborhood size is in Appendix D.5.

4.3 Where the degree-2 advantage comes from

In Section 4.2, g_2 is the strongest reader on average, and its margin is concentrated on near-OOD. Equation (13) splits g_2 into a mean term, which is g_1, a radial term, which is the whitened Mahalanobis score up to a batch-dependent sign, and a directional term. To see which part produces the near-OOD margin, Table 2 scores each term alone and g_2 with the directional term removed.

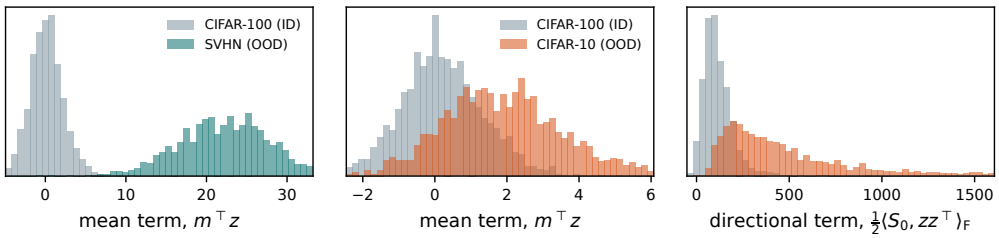
On far-OOD, g_1 alone achieves 0.954 AUROC, and g_2 improves on it only slightly. On near-OOD, the radial term alone stays at 0.744, close to g_1 (0.743), and adding it to the mean term gives only 0.755. The directional term alone instead attains 0.838, matching g_2 (0.836). The near-OOD gain of g_2 therefore comes from the directional term.

Figure 3 shows the two terms on held-out CIFAR-100 features. SVHN separates in the mean term, whereas CIFAR-10 is shifted but overlaps substantially with ID. In the directional term, ID scores stay small while a large part of CIFAR-10 falls in the right tail.

We next test whether the directional gain is a covariance effect rather than a mean shift, and whether the weighting of g_2 can be improved within the quadratic class. Here and below, the centered directional readout scores the directional term of Equation (13) alone, with S_0 replaced by the traceless

Table 2: Which term carries the near-OOD margin of g_2. Mean AUROC over far-OOD (10), near-OOD (8), and all (18) main-grid configurations (Equation (13); same 5-seed grid as Table 1).

Readout	What is scored	far-OOD	near-OOD	all
Mean term (g_1)	Mean shift	0.954	0.743	0.860
Radial term	Isotropic second moment (Mahalanobis)	0.809	0.744	0.780
Directional term	Anisotropic second moment	0.968	0.838	0.911
g_2 without directional term	Mean + radial	0.903	0.755	0.837
g_2	Mean + radial + directional	0.970	0.836	0.911



(a) Far-OOD.

(b) Near-OOD.

Figure 3: Degree-2 structure of q/p on the Table 1 protocol ($n = 2{,}000$, $\pi = 0.1$, ViT-B/16). Features are reference-whitened. The unlabeled batch supplies m and S_0 without labels; both terms are computed from the batch alone and evaluated on 2,000 held-out ID and 2,000 held-out OOD points. (a) CIFAR-100 vs. SVHN, mean term (teal: OOD; gray: ID). (b) The same protocol on CIFAR-10 (orange: OOD); left: mean term; right: directional term of Equation (13).

part C_0 of $\mathrm{Cov}_q = S + I - mm^\top$. Since $S_0 = C_0 + \left(mm^\top - \frac{\|m\|^2}{d} I\right)$, this removes the contribution of the batch mean. These controls are evaluated by five-fold cross-fitting within each batch (Appendix D.6). First, a mean shift alone also changes the covariance of q, adding extra variance along the mean-difference direction only. The centered directional readout could therefore still pick up the mean shift along this direction. We estimate the direction by m, remove it from C_0, and rescore. Near-OOD AUROC is essentially unchanged (Table 3), whereas the same projection reduces a synthetic mean-only shift to chance (Appendix D.7). Second, a ridge-regularized least-squares fit over the same quadratic class yields a near-OOD AUROC of 0.8305, compared with 0.8309 for g_2 (Appendix D.6).

Finally, far-OOD separation typically saturates with only a few leading eigen-directions of C_0, whereas near-OOD typically continues to benefit across tens of directions (Figure 4).

4.4 Finite-sample limits

Higher-degree readouts can express richer ratio structure, but the scale $n \asymp d^{k/2}$ of Section 3.3.3 grows quickly with the degree. At $n = 2{,}000$, this scale is below n at degree 2 ($d = 384$ or 768) but well above it at degree 3 (about 7,500 or 21,000). Consistent with this, the unregularized degree-3 plug-in does not improve AUROC over g_2 in any of the 18 real-data configurations (Appendix D.8).

Detecting a shift and recovering its sample ranking have different finite-sample costs. At $d = 64$, $\pi = 0.1$, $n = 10^4$, a degree-2 test detects the shift in every batch, whereas plug-in g_2 reaches AUROC 0.70 against 0.83 for the population g_2 (Table 10). Meeting the scale $n \asymp d^{k/2}$ is therefore not sufficient for ranking. The reference matters as well: on the Table 1 batches, near-OOD AUROC of g_2 rises from 0.738 at $n_{\mathrm{ref}} = 10^3$ to 0.853 at 10^4 and 0.860 at $5 \cdot 10^4$ (Appendix D.9).

Empirically, on ImageNet-1K at $\pi = 0.1$, the mean AUROC of g_2 exceeds that of Mahalanobis between $n = 300$ and $n = 500$ (Appendix D.4, Figure 5). Per pair, g_2 leads from $n \approx 500$ on four of five ViT-B/16 pairs and from $n = 300$ on three of five DINOv2 pairs, and on all ten pairs by $n = 2{,}000$. At $n = 500$ ($\pi = 0.1$), $\widehat{r}_{\mathrm{kNN}}$ inverts in 3 of 10 ImageNet-1K pairs and recovers when the batch-side neighborhood is shrunk (Appendix D.5).

Readout	near-OOD AUROC
g_1	0.7328
Centered directional	0.8299
Centered directional, m direction removed	0.8288
g_2	0.8309

Table 3: Cross-fitted near-OOD AUROC, mean over 8 configurations. The centered directional readout nearly matches g_2, and removing the mean direction barely changes it.

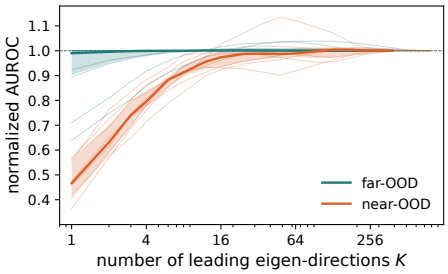


Figure 4: Normalized AUROC, $(\mathrm{AUROC}(K) - 0.5)/(\mathrm{AUROC}(\mathrm{all}) - 0.5)$, of the centered directional readout restricted to the K eigen-directions of C_0 with the largest absolute eigenvalues. Thin lines are Table 1 pairs; thick lines and bands give the median and interquartile range. Far-OOD is close to 1 from $K = 1$, whereas near-OOD reaches 0.9 at around $K = 8$.

4.5 Covariate-shifted ID and mixed OOD

A ratio reader also scores covariate-shifted ID in the batch as a departure from the reference. With half of the ImageNet-1K ID points taken from ImageNet-R, near-OOD AUROC of g_2 falls to 0.457, below 0.603 for Mahalanobis++ (Appendix D.10.1).

Splitting the OOD points over several sources also lowers the ratio readers, while the reference readers barely change (Appendix D.10.2).

5 Discussion

In the evaluated frozen representations, mean shifts account for much of the far-OOD separation. On near-OOD, the separation is instead carried by the directional term, whereas the radial term, which would capture an isotropic change in spread, adds little (Section 4.3). Near-OOD therefore appears less as a displacement of the batch than as a change in spread along particular directions, so that a weak mean shift can still coexist with useful separation. Near-OOD images may resemble ID classes closely enough to fall among the ID class clusters rather than away from the ID mean, while occupying some between-class directions more than the ID data do. Linderman et al. (2026) likewise find class-dependent covariance structure to be important. Although a reference reader can model anisotropic structure in the reference, it cannot tell along which directions the batch departs from the reference. Our results show that these directions can be estimated from the unlabeled batch without class labels.

The two regimes also differ in how many directions they require, with far-OOD separation typically saturating after a few leading eigen-directions of C_0 and near-OOD continuing to improve across tens of them (Figure 4). This would be expected if near-OOD occupies many between-class directions. Such a spread may partly explain why the near-OOD AUROC of g_2 is sensitive to the reference size (Section 4.4) and why splitting the OOD points over several sources lowers the ratio readers (Section 4.5).

6 Conclusion

An unlabeled test batch allows an OOD detector to score the batch-to-reference density ratio, but finite data limit how finely this ratio can be resolved. Detecting a shift can require fewer samples than recovering its sample ranking, and readout degree has to be chosen together with the available batch size. On the evaluated benchmarks, the closed-form degree-2 readout g_2 already matches or exceeds ratio estimators, with its near-OOD margin carried by the directions along which the batch covariance departs from the reference. A limitation is that the ratio readouts studied here do not distinguish covariate-shifted ID in the batch from OOD.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks that require disclosure: helping develop theoretical models and conceptual frameworks, formulating mathematical claims, providing critical ingredients for proofs, assisting in the writing of proofs, refining hypotheses, providing feedback on research methodology and experimental design, implementing methods, assisting with translation, and interpreting results. The underlying research concepts and ideas originated with the authors; suggested claims and proof ingredients were retained only after review. Generating synthetic datasets, cleaning or reformatting datasets, and qualitative data analysis are not applicable to this work.

Additionally, we used generative AI tools for tasks with recommended disclosure: creating and editing software code, drafting and editing parts of the research paper, searching for and summarizing literature, and brainstorming.

We have reviewed all AI-assisted work. All mathematical proofs were manually verified by the authors for correctness. All code was tested and validated by the authors. Literature references suggested by AI tools were checked against original sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Assumptions and complete proofs of the theoretical results in Section 3 are provided in Appendices A and B. The experimental protocol is specified in Section 4.1 and Appendix D.1, and detector definitions and hyperparameters are given in Appendix D.2. Anonymized code that runs the full pipeline and the raw results are included in the supplementary material.

REFERENCES

Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? Fixing ImageNet out-of-distribution detection evaluation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2471–2506, 2023. URL https://proceedings.mlr.press/v202/bitterwolf23a.html.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11(99):2973–3009, 2010. URL http://jmlr.org/papers/v11/blanchard10a.html.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pp. 93–104, 2000. ISBN 1581132174. doi: 10.1145/342009.335388. URL https://doi.org/10.1145/342009.335388.

Anthony L. Caterini and Gabriel Loaiza-Ganem. Entropic issues in likelihood-based OOD detection. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=ydqG4zPXo4w.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Xuefeng Du, Zhen Fang, Ilias Diakonikolas, and Yixuan Li. How does unlabeled data provably help out-of-distribution detection? In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jlEjB8MVGa.

Ke Fan, Tong Liu, Xingyu Qiu, Yikai Wang, Lian Huai, Zeyu Shangguan, Shuang Gou, Fengjian Liu, Yuqian Fu, Yanwei Fu, and Xingqun Jiang. Test-time linear out-of-distribution detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23752–23761, 2024. doi: 10.1109/CVPR52733.2024.02242.

Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=sde_7ZzGXOE.

Sebastian Farquhar and Yarin Gal. What 'out-of-distribution' is and is not. In *NeurIPS ML Safety Workshop*, 2022. URL https://openreview.net/forum?id=XCS_zBHQA2i.

Sahil Garg, Sanghamitra Dutta, Mina Dalirrooyfard, Anderson Schneider, and Yuriy Nevmyvaka. In- or out-of-distribution detection via dual divergence estimation. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 635–646, 2023. URL https://proceedings.mlr.press/v216/garg23b.html.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329, 2021. doi: 10.1109/ICCV48922.2021.00823.

Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*, 26 (2):309–336, 2011. doi: 10.1007/s10115-010-0283-2. URL https://doi.org/10.1007/s10115-010-0283-2.

Samuel Hopkins. *Statistical inference and the sum of squares method*. Cornell University, 2018.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461, 1946. ISSN 00804630. URL http://www.jstor.org/stable/97883.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(48):1391–1445, 2009. URL http://jmlr.org/papers/v10/kanamori09a.html.

Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training OOD detectors in their natural habitats. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10848–10865, 2022. URL https://proceedings.mlr.press/v162/katz-samuels22a.html.

Ao Ke, Wenlong Chen, Chuanwen Feng, Yukun Cao, Xike Xie, S. Kevin Zhou, and Lei Feng. Prototype-based optimal transport for out-of-distribution detection. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 3018–3026, 2025. doi: 10.24963/ijcai.2025/336. URL https://doi.org/10.24963/ijcai.2025/336. Main Track.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. ISSN 00034851. URL http://www.jstor.org/stable/2236703.

Dmitriy Kunisky, Alexander S. Wein, and Afonso S. Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. In *Mathematical Analysis, its Applications and Computation*, pp. 1–50, 2022. doi: 10.1007/978-3-030-97127-4_1.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge, 2015. CS 231N course project report.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Teresa Ledwina. Data-driven version of neyman's smooth test of fit. *Journal of the American Statistical Association*, 89(427):1000–1005, 1994.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Randolph Linderman, Noah Cowan, Yiran Chen, and Scott Linderman. A bayesian nonparametric perspective on mahalanobis distance for out of distribution detection. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=w3bMXPMDW1.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f5496252609c43eb8a3d147ab9b9c006-Paper.pdf.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJkXfE5xx.

Shuo Lu, Yingsheng Wang, Lijun Sheng, Lingxiao He, Aihua Zheng, and Jian Liang. Out-of-distribution detection: A task-oriented survey of recent advances. *ACM Computing Surveys*, 58 (2):1–39, 2025. doi: 10.1145/3760390.

Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7831–7840, 2022.

Maximilian Müller and Matthias Hein. Mahalanobis++: Improving OOD detection via feature normalization. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=vutMcZl50l.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Jerzy Neyman. "Smooth test" for goodness of fit. *Scandinavian Actuarial Journal*, 1937(3-4): 149–199, 1937.

Jerzy Neyman and Egon Sharpe Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231:289–337, 1933. doi: 10.1098/rsta.1933.0009. URL https://doi.org/10.1098/rsta.1933.0009.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

B. Picinbono and P. Duvaut. Geometrical properties of optimal volterra filters for signal detection. *IEEE Transactions on Information Theory*, 36(5):1061–1068, 1990. doi: 10.1109/18.57205.

Bernard Picinbono. On deflection as a performance criterion in detection. *IEEE Transactions on Aerospace and Electronic Systems*, 31(3):1072–1081, 1995.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021. URL https://arxiv.org/abs/2106.09022.

Oliver Rippel, Patrick Mertens, and Dorit Merhof. Modeling the distribution of normal data in pre-trained deep features for anomaly detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 6726–6733, 2021. doi: 10.1109/ICPR48806.2021.9412109.

Clayton Scott and Gilles Blanchard. Novelty detection: Unlabeled data definitely help. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 464–471, 2009. URL https://proceedings.mlr.press/v5/scott09a.html.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press Cambridge, 2012.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20827–20840, 2022. URL https://proceedings.mlr.press/v162/sun22d.html.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=5hLP5JY9S2d.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4911–4920, 2022.

Qing Wang, Sanjeev R. Kulkarni, and Sergio Verdu. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5): 2392–2405, 2009. doi: 10.1109/TIT.2009.2016060.

Jian Yan and Xianyang Zhang. Kernel two-sample tests in high dimensions: interplay between moment discrepancy and dimension-and-sample orders. *Biometrika*, 110(2):411–430, 2023.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022. URL https://arxiv.org/abs/2210.07242.

Yifeng Yang, Lin Zhu, Zewen Sun, Hengyu Liu, Qinying Gu, and Nanyang Ye. Oodd: Test-time out-of-distribution detection with dynamic dictionary. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30630–30639, 2025. doi: 10.1109/CVPR52734.2025.02852.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023. URL https://arxiv.org/abs/2306.09301.

Yi-Fan Zhang, Xue Wang, Tian Zhou, Kun Yuan, Zhang Zhang, Liang Wang, and Rong Jin. Model-free test time adaptation for out-of-distribution detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1542–1553, 2026. doi: 10.1109/TPAMI.2025.3615192.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009.

APPENDICES

A NOTATION AND STANDING ASSUMPTIONS

A.1 NOTATION

Symbol	Meaning
x, $z = \phi(x) \in \mathbb{R}^d$	Raw input, feature representation, and ambient dimension d.
p, a, q	Reference, anomaly, and test-batch densities with respect to a common dominating measure.
$\pi \in (0,1)$	Contamination rate, with $q = (1-\pi)p + \pi a$.
$r = q/p$	Batch-to-reference density ratio.
$\widehat{M}_\ell$	Empirical ℓth moment tensor from Definition 1.
V_k, ψ, G, b, β	Centered degree-$\leq k$ polynomial space, a basis, Gram matrix $\mathbb{E}_p[\psi\psi^\top]$, contrast $\mathbb{E}_q[\psi] - \mathbb{E}_p[\psi]$, and coefficients from Theorem 1.
$\mathcal{H}_\ell$, He_α	Exact-degree-ℓ Hermite shell and the multivariate probabilists' Hermite polynomial of multi-index α.
g_k, g_2	Degree-$\leq k$ Riesz representer and its closed form at $k = 2$.
$\widehat{m}$, $\widehat{S}$	Batch first moment and excess second moment.
n_{ref}, n	Reference and test-batch sample sizes.
$n_{\mathrm{OOD}} = \pi n$	Expected number of OOD samples in a batch of size n.
$\widehat{r}_{\mathrm{kNN}}$	Leave-one-out two-sample kNN density-ratio estimator.
k	Neighborhood size (for one-sample kNN or $\widehat{r}_{\mathrm{kNN}}$).
ΔH	Entropy gap $H(p) - H(a)$.
$J(a,p)$	Jeffreys divergence $\mathrm{KL}(a\|p) + \mathrm{KL}(p\|a)$.
$\widehat{G}_p$	Estimated mean gap $\widehat{\mathbb{E}}_a[s_p] - \widehat{\mathbb{E}}_p[s_p]$ of the canonical reference reader s_p.
$\mathrm{AUROC}(a,p;s)$	AUROC of score s, anomalous distribution first, with the midrank convention; see (14).

A.2 STANDING ASSUMPTIONS

1. **Dominating measure.** The distributions p, a, q have densities with respect to a common measure ν. The ratio $r = q/p$ is defined on $\mathrm{supp}(p)$; on $\{p = 0 < a\}$ we set $r = +\infty$.
2. **Finiteness.** Theorem 1 assumes finite moments of p through degree $2k$ and of q through degree k; Proposition 1 assumes finite displayed entropies and cross-entropies.
3. **Absolute continuity.** The condition $q \ll p$ is used to identify g_k with $\mathrm{Proj}_{V_k}(r-1)$.

4. **Sampling.** The reference is i.i.d. from p and the batch is i.i.d. from q, independently unless stated otherwise.

A.3 The whitening convention

Both samples are mapped according to

$$z \leftarrow \widehat{\Sigma}_{\mathrm{ref}}^{-1/2}\big(\phi(x) - \widehat{\mu}_{\mathrm{ref}}\big).$$

An invertible affine map preserves the mixture identity and density-ratio ordering. Whitening does not make the reference Gaussian: $p = \mathcal{N}(0, I_d)$ is used for the closed forms in Section 3.3.2; the general Gram form is given in Appendix B.3.

AUROC convention. Throughout,

$$\mathrm{AUROC}(a, p; s) := \Pr\{s(Z_a) > s(Z_p)\} + \tfrac{1}{2}\Pr\{s(Z_a) = s(Z_p)\}, \qquad Z_a \sim a, \quad Z_p \sim p, \tag{14}$$

so the anomalous distribution is listed first and ties receive midrank credit.

B Proofs and supplementary results for Section 3

B.1 The passage to AUROC and fitted reference densities

Proposition 2 (Calibration identity)**.** *Let F_p be the c.d.f. of $s(Z_p)$ and suppose it is continuous. Then*

$$\mathrm{AUROC}(a, p; s) - \tfrac{1}{2} = \mathbb{E}_a[F_p \circ s] - \mathbb{E}_p[F_p \circ s] = G(F_p \circ s).$$

Proof. Continuity of F_p rules out ties, so $\mathrm{AUROC} = P\big(s(Z_a) > s(Z_p)\big) = \mathbb{E}_a[F_p \circ s]$ by conditioning on Z_a, and $F_p(s(Z_p)) \sim \mathrm{Unif}[0, 1]$ by the probability integral transform, so $\mathbb{E}_p[F_p \circ s] = \frac{1}{2}$. □

Thus AUROC is the mean gap of the reference-calibrated score. Multiplying by π and using (6) gives (7).

Fitted reference densities. Because $q/p = (1-\pi) + \pi(a/p)$, the ratio in Proposition 1 is ranking-equivalent to the batch ratio. For a fitted reference density $\widetilde{p}$,

$$G(-\log \widetilde{p}) = \mathrm{KL}(a \| \widetilde{p}) - \mathrm{KL}(p \| \widetilde{p}) - \Delta H,$$

which reduces to Proposition 1 at $\widetilde{p} = p$. The extra $-\mathrm{KL}(p \| \widetilde{p})$ is a property of the fitted density, not of the estimand class, so two reference readers can take opposite sides of chance on the same (p, a).

B.2 Proof of Theorem 1

Choose a basis ψ of V_k and define

$$G = \mathbb{E}_p[\psi\psi^\top] \succ 0, \qquad b = \mathbb{E}_q[\psi] - \mathbb{E}_p[\psi],$$
$$\beta = G^{-1}b, \qquad g_k = \beta^\top \psi,$$

where $G \succ 0$ follows from the assumed non-degeneracy of $\langle \cdot, \cdot \rangle_p$ on V_k. Then $\langle f, g_k \rangle_p = \mathbb{E}_q[f] - \mathbb{E}_p[f]$ for every $f \in V_k$, so g_k is the Riesz representer of L. Deflection is the Rayleigh quotient

$$D(f) = \frac{\langle f, g_k \rangle_p}{\|f\|_p} \le \|g_k\|_p$$

by Cauchy–Schwarz, with equality if and only if $f = c g_k$ for $c > 0$; if $g_k = 0$ the numerator vanishes for every $f \in V_k$. This is (i). If $q \ll p$ and $r \in L^2(p)$, then $L(f) = \langle f, r - 1 \rangle_p$, so the projection theorem gives $g_k = \mathrm{Proj}_{V_k}(r - 1)$. This is (ii).

B.3 HERMITE FILTRATION AND THE GAUSSIAN CLOSED FORM

The working model $p = \mathcal{N}(0, I_d)$ names the complementary $L^2(p)$-orthogonal coordinate system on V_k. The probabilists' Hermite polynomials on $\mathbb{R}$ satisfy the generating function $\exp\big(xt - t^2/2\big) = \sum_{k\geq 0} \mathrm{He}_k(x)\, t^k/k!$, so $\mathrm{He}_0 = 1$, $\mathrm{He}_1(t) = t$, $\mathrm{He}_2(t) = t^2 - 1$, and $\mathbb{E}[\mathrm{He}_j(Z)\mathrm{He}_\ell(Z)] = j!\,\mathbf{1}_{j=\ell}$ for $Z \sim \mathcal{N}(0,1)$. For a multi-index $\alpha \in \mathbb{N}^d$ set

$$\mathrm{He}_\alpha(z) = \prod_{i=1}^{d} \mathrm{He}_{\alpha_i}(z_i).$$

Then $\{\mathrm{He}_\alpha\}$ is an orthogonal basis of $L^2(p)$ with $\|\mathrm{He}_\alpha\|_p^2 = \alpha!$. Write $\mathcal{H}_\ell = \mathrm{span}\{\mathrm{He}_\alpha : |\alpha| = \ell\}$ for the exact-degree-ℓ shell.

Lemma 1 (Hermite filtration)**.** *Under $p = \mathcal{N}(0, I_d)$, the sum $V_k = \bigoplus_{\ell=1}^{k} \mathcal{H}_\ell$ is $L^2(p)$-orthogonal, $\dim \mathcal{H}_\ell = \binom{d+\ell-1}{\ell}$, and $1 \perp V_k$. The representer of Theorem 1 therefore splits as $g_k = \sum_{\ell=1}^{k} g_k^{(\ell)}$ with $g_k^{(\ell)} \in \mathcal{H}_\ell$. If $q \ll p$ and $r - 1 \in L^2(p)$, then $g_k^{(\ell)} = \mathrm{Proj}_{\mathcal{H}_\ell}(r-1)$. In particular, an excess ratio $r - 1$ supported on a single shell $\mathcal{H}_\ell$ is invisible to every g_k with $k < \ell$, and is fully expressed by every g_k with $k \geq \ell$.*

Proof. Orthogonality of $\{\mathrm{He}_\alpha\}$ is the product of the one-dimensional identities. The set $\{\alpha : |\alpha| = \ell\}$ has cardinality $\binom{d+\ell-1}{\ell}$. Constants are $\mathcal{H}_0$, so centering removes them and $V_k = \bigoplus_{\ell=1}^{k} \mathcal{H}_\ell$. The Riesz representer therefore decomposes along this orthogonal sum. When $r - 1 \in L^2(p)$, Theorem 1(ii) identifies g_k with $\mathrm{Proj}_{V_k}(r-1)$, and the remaining claims are the corresponding statements for an orthogonal filtration. □

Derivation of g_2. Under $p = \mathcal{N}(0, I_d)$, the linear and quadratic monomials in V_2 are $\mathcal{H}_1$ and $\mathcal{H}_2$, which are $L^2(p)$-orthogonal by the lemma above. Write $f(z) = b^\top z + \langle A, zz^\top - I\rangle \in V_2$. Isserlis' formula is the Gram of this basis:

$$\mathbb{E}_p[(b^\top z)(c^\top z)] = b^\top c, \quad \mathbb{E}_p[\langle A, zz^\top - I\rangle\langle C, zz^\top - I\rangle] = 2\langle A, C\rangle_F,$$

and the linear–quadratic cross terms vanish. The representer equation gives $C = S/2$, hence

$$g_2(z) = m^\top z + \tfrac{1}{2}\langle S, zz^\top - I\rangle, \qquad \|g_2\|_p^2 = \|m\|^2 + \tfrac{1}{2}\|S\|_F^2.$$

The factor $\frac{1}{2}$ is the Gram weight of $\mathcal{H}_2$ in the Frobenius pairing.

Non-Gaussian Gram form. For any centered basis ϕ_2 of V_2 with nonsingular $\Gamma_{2,p}$, the representer is $g_{2,p} = \Delta_2^\top \Gamma_{2,p}^{-1}\phi_2$. Thus the Gaussian formula is a Gram surrogate rather than the general optimum. Within $\mathrm{span}\{m^\top z, \frac{1}{2}\langle S, zz^\top - I\rangle\}$, a 2×2 Gram solve gives the optimal combination in that span.

B.4 FINITE-RESOLUTION BLIND SPOTS

A fixed polynomial class cannot detect every alternative, even at the population level.

Corollary 1 (finite-resolution blind spots)**.** *Suppose there is a nonzero bounded $h \in L^2(p)$ with $\mathbb{E}_p h = 0$ and $h \perp V_k$. For any $0 < \epsilon < \|h\|_\infty^{-1}$, let $a_\epsilon = (1 + \epsilon h)p$ and $q = (1-\pi)p + \pi a_\epsilon$. Then $a_\epsilon \neq p$, but $g_k = 0$ and every $f \in V_k$ with $\mathrm{sd}_p(f) > 0$ has zero population deflection. Thus, no fixed V_k achieves positive deflection against every admissible alternative.*

Proof. The function a_ϵ is a density because $\mathbb{E}_p h = 0$ and the bound on ϵ ensures positivity. Moreover, $q/p - 1 = \pi\epsilon h$, whose projection onto V_k is zero. The claim follows from Theorem 1. □

Positive deflection does not ensure useful ranking. Even a nonzero population g_2 can have arbitrarily low AUROC. Let $p = \mathcal{N}(0,1)$ and

$$a = (1-\epsilon)\mathcal{N}(0, \sigma^2) + \frac{\epsilon}{2}\mathcal{N}(L, \sigma^2) + \frac{\epsilon}{2}\mathcal{N}(-L, \sigma^2),$$

where $0 < \epsilon < 1$, $L > 0$, and $\epsilon L^2 > 1$. For $q = (1-\pi)p + \pi a$ with $\pi > 0$, Equation (11) gives $g_2(z) = \frac{\pi}{2}(\sigma^2 + \epsilon L^2 - 1)(z^2 - 1)$. Its mean gap is positive, yet, by conditioning on the mixture component,

$$\lim_{\sigma \downarrow 0} \mathrm{AUROC}(a, p; g_2) = \epsilon \Pr_{Z \sim p}(|Z| < L) \leq \epsilon.$$

Choosing ϵ arbitrarily small and $L > \epsilon^{-1/2}$ therefore yields arbitrarily poor ranking. This is a population approximation failure, not a finite-sample effect.

Neighborhood readouts. Neighborhood readouts lie outside every finite-degree class: for every K, there exist distributions with identical moments through degree K but different local densities. The kNN ratio estimator used in the experiments is defined in Appendix D.2.

B.5 Sample cost of shell detection

By Lemma 1, the degree-k Hermite shell has dimension $M_k = \binom{d+k-1}{k} = \Theta(d^k)$. Let $h_k(z) = (\mathrm{He}_\alpha(z)/\sqrt{\alpha!})_{|\alpha|=k}$, $\theta_k = \mathbb{E}_q h_k(Z)$, and $\rho^2 = \|\theta_k\|^2$ (the population signal energy, a deterministic quantity). For i.i.d. $Z_1, \ldots, Z_n \sim q$, define the empirical counterparts

$$\widehat{\theta}_k = \frac{1}{n}\sum_{i=1}^{n} h_k(Z_i), \qquad \widehat{T}_k = \|\widehat{\theta}_k\|^2.$$

Note that $\widehat{T}_k$ is a biased estimator of ρ^2: even under the null ($q = p$), its expectation is $M_k/n > 0$. Under $q = p = \mathcal{N}(0, I_d)$, orthonormality gives $\mathbb{E}_p h_k = 0$ and $\mathbb{E}_p[h_k h_k^\top] = I_{M_k}$. Consequently,

$$\mathbb{E}_0 \widehat{T}_k = \frac{M_k}{n}, \qquad \mathrm{Var}_0(\widehat{T}_k) = \frac{2(n-1)M_k}{n^3} + \frac{\mathrm{Var}_p(\|h_k(Z)\|^2)}{n^3}.$$

Indeed, the diagonal sum and the degenerate off-diagonal sum in $n^2\widehat{T}_k$ are uncorrelated under the null. Thus $\mathrm{Var}_0(\widehat{T}_k) \sim 2M_k/n^2$ when $n \to \infty$ and $\mathrm{Var}_p(\|h_k(Z)\|^2) = o(nM_k)$; this condition must also be checked if dimension or degree grows. The distinct, diagonal-debiased statistic

$$U_k = \frac{1}{n(n-1)} \sum_{i \neq j} h_k(Z_i)^\top h_k(Z_j)$$

has $\mathbb{E}_q U_k = \rho^2$ and exact null variance $2M_k/[n(n-1)]$ for $n > 1$. Comparing this signal with the null standard deviation gives the heuristic activation scale

$$n \asymp \frac{\sqrt{M_k}}{\rho^2} = \Theta(d^{k/2})$$

at fixed energy. This is a null-calibrated energy-detection scale, not a general power guarantee: variance under the alternative also matters. It does not establish coefficient-direction recovery. For example, whenever the coefficient covariance exists, $\mathbb{E}\|\widehat{\theta}_k - \theta_k\|^2 = \mathrm{tr}(\mathrm{Cov}_q\, h_k)/n$, which is M_k/n under the null. Related rigorous testing asymptotics are given by Yan & Zhang (2023).

C Moment updates and leave-one-out scoring

Remark 3 (A stream is a growing batch). The moments in Definition 1 update without retaining raw samples. At degree 2, the state $(n, \widehat{m}, \widehat{S})$ updates in $O(d^2)$ per arrival, as detailed below.

Equation (11) depends on the test sample only through $m = \mathbb{E}_q[z]$ and $S = \mathbb{E}_q[zz^\top] - I$. The empirical moments

$$\widehat{m}_n = n^{-1}\sum_{i=1}^{n} z_i, \qquad \widehat{S}_n = n^{-1}\sum_{i=1}^{n} z_i z_i^\top - I$$

are additive sufficient statistics for every degree-2 readout. A new whitened point z_{n+1} updates them by

$$\widehat{m}_{n+1} = \frac{n}{n+1}\widehat{m}_n + \frac{1}{n+1} z_{n+1}, \qquad \widehat{S}_{n+1} = \frac{n}{n+1}(\widehat{S}_n + I) + \frac{1}{n+1} z_{n+1} z_{n+1}^\top - I. \quad (15)$$

The running state is $(n, \widehat{m}, \widehat{S})$; the raw points need not be stored. The reference whitening of Appendix A.3 is computed once and applied to each arrival.

The leave-one-out scores used in the experiments are the same statistics after a rank-one downdate,

$$\widehat{m}^{(-)} = \frac{n\widehat{m} - z}{n-1}, \qquad \widehat{S}^{(-)} = \frac{n(\widehat{S} + I) - zz^\top}{n-1} - I.$$

To score an arriving point without the leave-in self-alignment term described below, evaluate $\widehat{g}_2$ at the pre-update moments and then apply Equation (15).

Leave-in bias and the necessity of leave-one-out scoring. A leave-in estimate can separate symmetric populations even when the population mean shift m is zero. For $\widehat{m}$ one has

$$\widehat{m}^\top z_j = n^{-1}\|z_j\|^2 + n^{-1}\sum_{i \neq j} z_i^\top z_j,$$

and the self-alignment term $n^{-1}\|z_j\|^2$ is nonzero even when the population mean shift m is zero. If instead $\widehat{m}$ is fitted independently of the query and both p and a are centrally symmetric, each law of $\widehat{m}^\top Z$ is symmetric conditionally on $\widehat{m}$, so the midrank AUROC is exactly $1/2$.

The main g_1 and g_2 readouts therefore use the LOO downdate above to remove direct query self-alignment.

D Experimental details and additional results

D.1 Protocol details

Unless stated otherwise, experiments use $n_{\text{ref}} = 5{,}000$, $n = 2{,}000$, and $\pi = 0.1$; each subsection states any deviation. The ID and OOD datasets follow the OpenOOD v1.5 benchmark (Yang et al., 2022; Zhang et al., 2023). The CIFAR-10 and CIFAR-100 (Krizhevsky, 2009) grids use the other CIFAR dataset as near-OOD and SVHN (Netzer et al., 2011) as far-OOD. The ImageNet-1K (Deng et al., 2009) grid uses SSB-hard (Vaze et al., 2022) and NINCO (Bitterwolf et al., 2023) as near-OOD, and iNaturalist (Van Horn et al., 2018), Textures (Cimpoi et al., 2014), and OpenImage-O (Wang et al., 2022) as far-OOD.

The primary ViT-B/16 (Dosovitskiy et al., 2021) features are $d = 768$ from torchvision `IMAGENET1K_V1`; DINOv2 ViT-S/14 (Oquab et al., 2024) features are $d = 384$ from `facebookresearch/dinov2` (`dinov2_vits14`). Features are the CLS token after the final LayerNorm (the input to the classification head); images are resized to 224×224 and normalized with the ImageNet mean and standard deviation.

Each real-data configuration is the mean over five seeds (0–4). Each seed resamples the reference bank, test ID points, and test OOD points once.

D.2 Detectors and external baselines

This section defines all detectors and baselines used in Section 4.

Detectors.

- g_1: leave-one-out batch mean, $\widehat{m}_{-i}^\top z_i$, on reference-whitened features z (Appendix A.3; numerical ridge $10^{-5} I_d$).
- g_2: leave-one-out closed form $\widehat{m}_{-i}^\top z_i + \frac{1}{2}\langle \widehat{S}_{-i},\, z_i z_i^\top - I \rangle$, untruncated, on the same whitened features. This is the finite-sample plug-in of Theorem 1.
- $\widehat{r}_{\text{kNN}}$: leave-one-out two-sample kNN ratio, scored by the radius ratio $R_{p,k}(z)/R_{q,k}(z)$, where $R_{p,k}$ and $R_{q,k}$ are the distances to the kth nearest neighbor in the reference and batch banks; $k = 50$, ℓ_2-normalized features (Sun et al., 2022). This is a monotone transform of the kNN density-ratio estimate (Wang et al., 2009), which is consistent for $k \to \infty$ and $k/N \to 0$, with N the bank size.

- Mahalanobis: label-free single-Gaussian marginal score on the unlabeled reference bank, on the same whitened features.
- One-sample kNN: distance to the kth reference neighbor, with Sun's reference-bank-scaled neighborhoods: $k = 5$ on CIFAR-10, $k = 20$ on CIFAR-100, and $k = 4$ on ImageNet-1K/200; ℓ_2-normalized (Sun et al., 2022).
- Mahalanobis++ (Müller & Hein, 2025): nearest class-conditional Mahalanobis score on ℓ_2-normalized features, using ID class labels on the reference bank only.

Additional baselines.

- uLSIF (Kanamori et al., 2009): RBF centers at all test-batch points. Five-fold cross-validation on the uLSIF risk selects $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and a bandwidth multiplier in $\{0.6, 0.8, 1.0, 1.2, 1.5, 2.0\}$ around the median heuristic. The estimated ratio is clipped at zero and the coefficients are left unclipped. Scores are leave-in.
- MMD witness (Gretton et al., 2012): RBF kernel with bandwidth set by the median heuristic on up to $1{,}000$ reference and $1{,}000$ test points. The score is the batch-minus-reference kernel witness and is leave-in.
- C2ST-LR (Lopez-Paz & Oquab, 2017): standardized linear logistic regression with $C = 1$, evaluated by two-fold cross-fitting with a 1:1 split of each bank.
- LOF (Breunig et al., 2000): $k = 50$, following the baseline configuration of Sun et al. (2022).
- OODD (Yang et al., 2025): a random 50% of the reference as the ID dictionary, the batch streamed in chunks of 512 with the official queue lengths (128/512/2048) and $K_2 = 5$; K_1 is the official value scaled by $n_{\text{ref}}/N_{\text{train}}$ as for one-sample kNN, which gives $K_1 = 1$ on every configuration except ImageNet-200, where $K_1 = 2$.
- RTL++ (Fan et al., 2024): the default settings (60% of the batch kept, Lasso $\alpha = 10^{-5}$) on the energy of a linear cross-entropy probe fitted on the labeled reference bank.
- RMDS (Ren et al., 2021): labeled class-conditional Mahalanobis score relative to a background model; reference-only, but not a monotone transform of p. AUROC is reported in Table 5.
- DDE (Garg et al., 2023): one critic per configuration, using cached encoder features and the official implementation (commit `ebc1d81`). The critic is a three-layer width-64 MLP with no dropout, trained with `std`=0.15, learning rate 10^{-4}, batch size $10{,}000$, and $2{,}450$ iterations; its raw dual output is the score.

D.3 Complete tables

Tables 4–6 report FPR@95 and additional baselines.

At the FPR@95 operating point, g_2 is at least as good as both Mahalanobis and one-sample kNN in 14 of 18 configurations.

Tables 7 and 8 summarize the full OpenOOD v1.5 grid, which adds TIN (Le & Yang, 2015), MNIST (Lecun et al., 1998), Textures, and Places365 (Zhou et al., 2018) for CIFAR and ImageNet-200 as a fourth ID dataset (44 configurations), under the same protocol as Table 1. On this grid, g_2 has the highest mean AUROC and the lowest mean FPR@95 overall, on near-OOD, and on far-OOD. On ImageNet-200, Mahalanobis++ has a lower mean FPR@95 than g_2 (0.265 vs 0.267), and on its near-OOD sets RMDS (0.885) and Mahalanobis++ (0.881) match g_2 (0.881) in AUROC.

D.4 Batch size and contamination sweeps

Sweep details. On the ImageNet-1K sweep configurations, g_2 stays ahead of Mahalanobis from $n \approx 500$ on four of five ViT-B/16 pairs (all five by $n = 1{,}000$), and from $n = 300$ on three of five DINOv2 pairs (all five by $n = 2{,}000$). Mahalanobis leads in 57 of 120 configurations, 39 at $n \leq 100$. At fixed $n = 2{,}000$, g_2 outperforms Mahalanobis at $\pi = 0.05$ in 8 of 10 configurations.

Table 4: FPR@95 on the Table 1 configurations ($\pi = 0.1$, $n = 2{,}000$). Lower is better; chance at this operating point is 0.95. Bold marks the lowest value in each row. Column order matches Table 1.

				reference density p			density ratio q/p			
ref.	encoder	OOD	type	Maha.	Maha++	kNN	g_1	g_2	$\widehat{r}_{\mathrm{kNN}}$	DDE
CIFAR-10	ViT-B/16	SVHN	far	0.346	0.401	0.198	**0.000**	**0.000**	**0.000**	**0.000**
		CIFAR-100	near	0.608	0.534	**0.415**	0.844	0.571	0.432	0.537
	DINOv2	SVHN	far	0.527	0.222	0.315	0.001	0.001	**0.000**	**0.000**
		CIFAR-100	near	0.743	0.624	0.650	0.829	0.682	**0.453**	0.805
CIFAR-100	ViT-B/16	SVHN	far	0.898	0.916	0.597	**0.000**	**0.000**	**0.000**	**0.000**
		CIFAR-10	near	0.747	0.686	0.813	0.737	**0.453**	0.485	0.628
	DINOv2	SVHN	far	0.932	0.676	0.738	0.002	0.002	**0.000**	**0.000**
		CIFAR-10	near	0.913	0.784	0.864	0.771	0.676	**0.529**	0.842
ImageNet-1K	ViT-B/16	iNaturalist	far	0.339	0.192	0.446	0.009	**0.006**	0.010	0.010
		Textures	far	0.617	0.424	0.519	0.342	**0.256**	0.635	0.309
		OpenImage-O	far	0.405	**0.319**	0.523	0.614	0.335	0.834	0.360
		SSB-hard	near	0.873	0.839	0.891	0.895	**0.830**	0.962	0.856
		NINCO	near	0.555	0.460	0.679	0.688	**0.370**	0.932	0.499
	DINOv2	iNaturalist	far	0.060	0.039	0.141	0.014	**0.007**	0.015	0.032
		Textures	far	0.415	0.465	0.599	0.282	**0.120**	0.266	0.491
		OpenImage-O	far	0.665	**0.360**	0.419	0.794	0.655	0.584	0.796
		SSB-hard	near	**0.794**	0.839	0.857	0.951	0.815	0.831	0.914
		NINCO	near	0.786	0.643	0.716	0.824	**0.591**	0.774	0.790
Mean, far-OOD (10)				0.520	0.401	0.450	0.206	**0.138**	0.234	0.200
Mean, near-OOD (8)				0.752	0.676	0.736	0.818	**0.623**	0.675	0.734
Mean, all (18)				0.623	0.524	0.577	0.478	**0.354**	0.430	0.437

D.5 Neighborhood size of kNN readers

One-sample kNN (reference reader). On ImageNet-1K with ViT-B/16 ($n = 2{,}000$, $\pi = 0.1$, 5 seeds), one-sample kNN is evaluated for $k \in \{1, 4, 5, 20, 50, 200, 1000\}$, $n_{\mathrm{ref}} \in \{5{,}000,\ 50{,}000\}$, and with and without reference whitening. On every OOD source, g_2 exceeds the best one-sample kNN AUROC over this grid, by 0.068 on average and by at least 0.035.

kNN ratio (ratio reader). At $n = 500$, $\pi = 0.1$, and $k = 50$, $\widehat{r}_{\mathrm{kNN}}$ inverts in 3 of 10 ImageNet-1K encoder–OOD pairs, all with ViT-B/16. To examine the role of the neighborhood size, the reference-side k_{ref} is fixed at 50 and the batch-side k is varied. Reducing the batch-side k to 5 restores all three pairs: SSB-hard rises from 0.424 to 0.653, NINCO from 0.294 to 0.739, and OpenImage-O from 0.376 to 0.784. A small k is not uniformly better, however. At $n = 2{,}000$, increasing k from 50 to 200 lowers the pooled AUROC of these three pairs from 0.638 to 0.349, below chance, while $k = 2$–5 lowers DINOv2/OpenImage-O to 0.380–0.436. The suitable k therefore depends on the batch size and the encoder.

D.6 Cross-fitted quadratic comparison

This section compares the Gaussian-working-model weights of g_2 with quadratic weights fitted to the observed reference. All readouts share a label-free five-fold split of each Table 1 configuration ($n = 2{,}000$, $\pi = 0.1$, 5 seeds): parameters are fitted on the $1{,}600$ batch points outside the held-out fold and scored on that fold; AUROC is computed within fold. Paired 95% CIs are hierarchical bootstraps ($5{,}000$ resamples, seeds paired within configuration).

The quadratic least-squares estimator minimizes the two-sample LS risk over the full Hermite quadratic class ($74{,}304$ features at $d = 384$, $296{,}064$ at $d = 768$) in exact dual form, with the ridge selected by nested two-sample validation (reference $4{,}000/1{,}000$, batch $1{,}280/320$); no ID/OOD label enters selection. The selected ridge gives near-OOD AUROC 0.8305, compared with 0.8309

Table 5: Additional baselines on the Table 1 configurations ($\pi = 0.1$, $n = 2{,}000$). AUROC; bold is best in the row. FPR@95 for the same columns is Table 6.

				reference-only		test-time adapters		density-ratio / two-sample			
ref.	encoder	OOD	type	LOF	RMDS	OODD	RTL++	uLSIF	MMD	C2ST	g_2
CIFAR-10	ViT-B/16	SVHN	far	0.653	0.944	0.999	0.965	**1.000**	**1.000**	0.974	**1.000**
		CIFAR-100	near	0.843	**0.918**	0.856	0.901	0.851	0.799	0.648	0.880
	DINOv2	SVHN	far	0.776	0.979	0.999	0.993	**1.000**	**1.000**	0.994	**1.000**
		CIFAR-100	near	0.867	**0.914**	0.834	0.898	0.746	0.758	0.686	0.855
CIFAR-100	ViT-B/16	SVHN	far	0.525	0.862	0.997	0.953	**1.000**	0.999	0.961	**1.000**
		CIFAR-10	near	0.670	0.820	0.760	0.780	0.744	0.810	0.691	**0.902**
	DINOv2	SVHN	far	0.549	0.925	0.997	0.960	**1.000**	**1.000**	0.987	**1.000**
		CIFAR-10	near	0.722	0.818	0.805	0.685	0.733	0.761	0.683	**0.843**
ImageNet-1K	ViT-B/16	iNaturalist	far	0.830	0.936	0.965	0.827	0.995	0.993	0.903	**0.996**
		Textures	far	0.830	0.872	0.887	0.819	**0.958**	0.923	0.809	0.956
		OpenImage-O	far	0.842	0.903	0.885	0.790	0.892	0.927	0.741	**0.928**
		SSB-hard	near	0.581	0.674	0.667	0.581	0.716	0.669	0.604	**0.728**
		NINCO	near	0.755	0.857	0.834	0.721	0.875	0.858	0.701	**0.912**
	DINOv2	iNaturalist	far	0.932	0.971	0.934	0.898	0.994	0.989	0.929	**0.996**
		Textures	far	0.786	0.857	0.845	0.795	0.970	0.959	0.860	**0.976**
		OpenImage-O	far	0.765	**0.904**	0.764	0.795	0.767	0.754	0.668	0.851
		SSB-hard	near	0.697	0.652	0.589	0.556	0.529	0.622	0.535	**0.716**
		NINCO	near	0.606	0.801	0.759	0.691	0.718	0.757	0.642	**0.857**

Table 6: FPR@95 on the Table 5 configurations ($\pi = 0.1$, $n = 2{,}000$). Lower is better; chance at this operating point is 0.95. Bold is lowest in the row. Column order matches Table 5.

				reference-only		test-time adapters		density-ratio / two-sample			
ref.	encoder	OOD	type	LOF	RMDS	OODD	RTL++	uLSIF	MMD	C2ST	g_2
CIFAR-10	ViT-B/16	SVHN	far	0.707	0.172	0.001	0.133	**0.000**	**0.000**	0.145	**0.000**
		CIFAR-100	near	0.570	**0.284**	0.549	0.445	0.641	0.677	0.935	0.571
	DINOv2	SVHN	far	0.741	0.075	0.001	0.024	**0.000**	**0.000**	0.028	0.001
		CIFAR-100	near	0.659	**0.316**	0.616	0.380	0.773	0.760	0.882	0.682
CIFAR-100	ViT-B/16	SVHN	far	0.754	0.317	0.001	0.181	**0.000**	0.001	0.171	**0.000**
		CIFAR-10	near	0.940	0.588	0.745	0.671	0.704	0.588	0.851	**0.453**
	DINOv2	SVHN	far	0.935	0.228	0.006	0.144	**0.000**	**0.000**	0.062	0.002
		CIFAR-10	near	0.847	**0.548**	0.596	0.755	0.695	0.632	0.872	0.676
ImageNet-1K	ViT-B/16	iNaturalist	far	0.485	0.275	0.206	0.542	0.008	0.012	0.553	**0.006**
		Textures	far	0.668	0.470	0.489	0.602	**0.247**	0.486	0.743	0.256
		OpenImage-O	far	0.501	0.366	0.391	0.691	0.531	0.338	0.871	**0.335**
		SSB-hard	near	0.880	0.847	0.850	0.903	**0.792**	0.806	0.928	0.830
		NINCO	near	0.666	0.487	0.519	0.797	0.503	0.479	0.876	**0.370**
	DINOv2	iNaturalist	far	0.271	0.120	0.368	0.315	0.015	0.046	0.376	**0.007**
		Textures	far	0.615	0.627	0.658	0.646	0.174	0.262	0.625	**0.120**
		OpenImage-O	far	0.848	**0.437**	0.670	0.587	0.741	0.734	0.884	0.655
		SSB-hard	near	**0.801**	0.874	0.885	0.917	0.914	0.897	0.950	0.815
		NINCO	near	0.904	0.667	0.721	0.831	0.737	0.740	0.879	**0.591**

for g_2. Along the path $(\widehat{G} + \lambda I)^{-1}\widehat{b}$, with $\widehat{G}$ the reference Gram of these features and $\widehat{b}$ the batch-minus-reference contrast, near-OOD AUROC is 0.8203, 0.8209, 0.8241, 0.8291, 0.8307, 0.8309 for $\lambda \in \{10^{-3}, 1, 10, 10^2, 10^3, 10^4\}$, and 0.8309 with the Gaussian Gram, which coincides with g_2 to within 10^{-4}.

Table 7: Mean AUROC on the full OpenOOD v1.5 grid by ID dataset and OOD type (44 configurations; same protocol as Table 1). Bold marks the best value in each column.

	CIFAR-10 (12)	CIFAR-100 (12)	ImageNet-200 (10)	ImageNet-1K (10)	far (28)	near (16)	all (44)
reference-only							
Mahalanobis	0.950	0.818	0.904	0.825	0.903	0.826	0.875
Mahalanobis++	0.961	0.845	0.931	0.858	0.921	0.861	0.899
One-sample kNN	0.959	0.843	0.923	0.792	0.914	0.824	0.881
LOF	0.912	0.782	0.850	0.762	0.852	0.787	0.828
RMDS	0.955	0.874	0.930	0.843	0.927	0.856	0.901
test-time adapters							
OODD	0.966	0.896	0.841	0.813	0.938	0.789	0.884
RTL++	0.966	0.851	0.894	0.747	0.911	0.794	0.869
density-ratio / two-sample							
g_1	0.964	0.965	0.876	0.833	0.972	0.814	0.915
g_2	**0.977**	**0.977**	**0.939**	**0.892**	**0.985**	**0.887**	**0.949**
$\widehat{r}_{\mathrm{kNN}}$	0.974	0.950	0.868	0.802	0.966	0.796	0.904
DDE	0.972	0.964	0.905	0.855	0.975	0.846	0.928
uLSIF	0.966	0.956	0.843	0.841	0.968	0.801	0.907
MMD	0.962	0.960	0.869	0.845	0.970	0.815	0.914
C2ST	0.908	0.904	0.758	0.739	0.907	0.708	0.834

Table 8: Mean FPR@95 on the full OpenOOD v1.5 grid by ID dataset and OOD type (44 configurations; same protocol as Table 1). Lower is better; bold marks the lowest value in each column. Column and row order match Table 7.

	CIFAR-10 (12)	CIFAR-100 (12)	ImageNet-200 (10)	ImageNet-1K (10)	far (28)	near (16)	all (44)
reference-only							
Mahalanobis	0.202	0.413	0.332	0.551	0.266	0.547	0.368
Mahalanobis++	0.159	0.366	**0.265**	0.458	0.211	0.476	0.307
One-sample kNN	0.151	0.365	0.304	0.579	0.234	0.529	0.341
LOF	0.280	0.524	0.540	0.664	0.403	0.651	0.493
RMDS	0.160	0.398	0.277	0.517	0.245	0.485	0.333
test-time adapters							
OODD	0.135	0.372	0.527	0.576	0.260	0.614	0.389
RTL++	0.140	0.458	0.387	0.683	0.293	0.604	0.406
density-ratio / two-sample							
g_1	0.140	0.127	0.489	0.541	0.131	0.615	0.307
g_2	**0.106**	**0.099**	0.267	**0.398**	**0.071**	**0.446**	**0.207**
$\widehat{r}_{\mathrm{kNN}}$	**0.106**	0.211	0.493	0.584	0.163	0.625	0.331
DDE	0.112	0.125	0.379	0.506	0.113	0.532	0.266
uLSIF	0.118	0.118	0.421	0.466	0.113	0.534	0.266
MMD	0.125	0.122	0.413	0.480	0.120	0.534	0.270
C2ST	0.365	0.368	0.792	0.769	0.399	0.827	0.555

D.7 Origin of the covariance signal

Under the mixture, Cov_q contains the rank-one term $\pi(1-\pi)(\mu_{\mathrm{OOD}}-\mu_{\mathrm{ID}})(\mu_{\mathrm{OOD}}-\mu_{\mathrm{ID}})^\top$, so a mean difference alone changes the covariance. On each fit of Appendix D.6, u is the unit estimated mean-difference direction, $H = I - uu^\top$, and the projected readout scores with $HC_0H - \mathrm{tr}(HC_0H)H/(d-1)$, so the removed dimension does not leak a radial term. Near-OOD AUROC (Table 3): centered directional 0.8299; mean direction removed 0.8288 ($+0.0959$ $[+0.0781, +0.1142]$ over g_1); a Haar-random direction removed 0.8299. In a synthetic control with $p = \mathcal{N}(0, I)$ and mean shift $1.5e_1$, the covariance-only oracle AUROC is 0.752 before projecting out e_1 and exactly 0.500 after, while a covariance shift $I + 0.8e_2e_2^\top$ survives the projection (0.593).

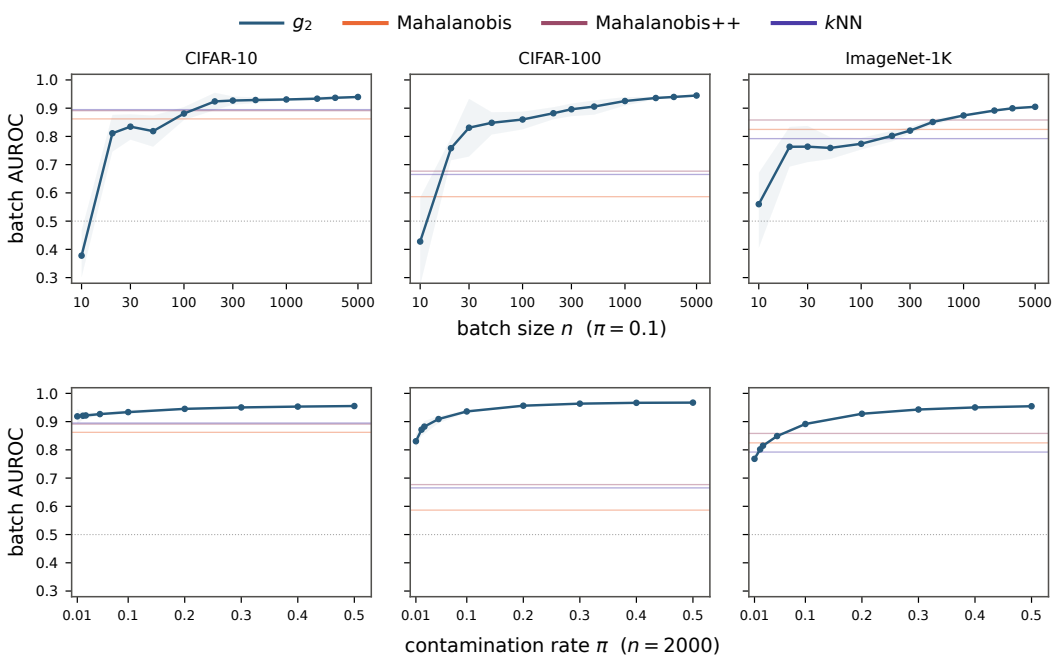


Figure 5: n and π sweeps of batch AUROC on CIFAR-10, CIFAR-100, and ImageNet-1K. Top: batch size at $\pi = 0.1$. Bottom: contamination rate at $n = 2{,}000$. AUROC is computed per setting, then averaged equally over OOD sources and the two encoders, then over five seeds. Bands are pointwise 95% bootstrap intervals for that mean, from resampling seeds with encoders and OOD sources held fixed. Mahalanobis, Mahalanobis++, and one-sample kNN do not use the test batch.

D.8 Higher-degree readouts and estimation cost

This section supports the cost-of-resolution analysis of Section 3.3.3.

D.8.1 Real data

Setting. Same configurations as Table 1; the degree-3 readout adds $\frac{1}{3!}$ times the leave-one-out third-order Hermite shell to g_2.

Result. The degree-3 plug-in does not improve on g_2 in any of the 18 configurations. The mean absolute AUROC difference is 0.039; the maximum deviation is a loss, not a gain.

D.8.2 Synthetic data

Setting. The reference is $p = \mathcal{N}(0, I_d)$. Along a random unit direction, the alternatives are $\mathcal{N}(3, 1)$, $\mathcal{N}(0, 4)$, and a Gamma distribution with shape $4/9$ standardized to mean zero and variance one, so that they first differ from p in a moment of degree 1, 2, and 3, respectively. The oracle/plug-in comparison (Table 9) replaces the one-direction variance change with a rank-4 subspace of variance 4 ($d = 16$). The mixture rate is $\pi = 0.1$, over 20 seeds.

Oracle versus plug-in readouts. Table 9 scores each readout once with population moments (oracle) and once with batch moments (plug-in) on points independent of the batch, separating approximation error from estimation error. Plug-in g_2 on the subspace alternative rises from 0.682 at $n = 10^2$ to 0.891 at $n = 10^4$ ($d = 16$, rank 4); plug-in g_3 on the skewed alternative stays within 0.06 of chance at $d = 16$ and within 0.011 at $d = 64$ for every $n \leq 10^4$.

Sample cost of detection. The exponents of the sample cost are fitted over five dimensions, $d \in \{8, 16, 32, 64, 128\}$, using median standardized energy $z > 3$ as the activation criterion. The fitted

Table 9: Synthetic alternatives: population (oracle) versus plug-in readouts, AUROC on points independent of the batch ($\pi = 0.1$, 20 seeds). $A^\star$ is the Neyman–Pearson optimum. Plug-in values at $n = 10^4$. The order is the first differing moment order; plug-in g_3 is shown for the skewed alternatives.

alternative	order	$A^\star$	oracle g_1	oracle g_2	oracle g_3	plug-in g_2	plug-in g_3
mean shift ($d = 16$)	1	0.983	0.983	0.978	0.955	0.978	—
subspace variance ($d = 16$, rank 4)	2	0.897	0.500	0.897	0.897	0.891	—
skew ($d = 16$)	3	0.826	0.500	0.500	0.636	0.499	0.553
skew ($d = 64$)	3	0.824	0.500	0.500	0.636	0.502	0.511

exponents (SE) are 0.566 ± 0.062, 1.029 ± 0.080, and 1.412 ± 0.068 for degrees 1–3. This experiment concerns the detection statistic T_k, not labeled AUROC or estimation of the ranking direction.

Detection versus ranking on the same batch. For the joint mean-and-covariance alternative at $d = 64$, $\pi = 0.1$, 100 batches, and nested prefixes $n \in \{10^2, \ldots, 10^4\}$, each batch is scored for (i) rejection of the unbiased degree-2 energy statistic at the 95th percentile of 5,000 null batches, (ii) the cosine between estimated and population Hermite coefficients of g_2, and (iii) AUROC of plug-in g_2 on $2{,}000 + 2{,}000$ independent points (Table 10). Population g_2 attains AUROC 0.832 and the true ratio 0.848 at every n. The gap between 0.832 and plug-in g_2 in Table 10 is estimation error, and the remaining gap of 0.015 up to 0.848 is the error from truncating at degree 2.

Table 10: Detection power (Wilson 95%), coefficient cosine, and AUROC on the same synthetic batches ($d = 64$, $\pi = 0.1$, 100 repeats).

n	power	cosine	plug-in g_2
10^2	0.01 [0.002, 0.054]	0.048	0.527
10^3	0.22 [0.150, 0.311]	0.150	0.584
$3 \cdot 10^3$	0.73 [0.636, 0.807]	0.256	0.635
10^4	1.00 [0.963, 1.000]	0.438	0.705

D.9 Reference size and low contamination

We score the Table 1 batches themselves with g_1, g_2, Mahalanobis, Mahalanobis++, and one-sample kNN against references with n_{ref} from 10^3 to $5 \cdot 10^4$, scaling the neighborhood of one-sample kNN (Sun et al., 2022) in proportion to n_{ref}. On average, all readers improve with n_{ref}, and near-OOD AUROC of g_2 levels off above $n_{\mathrm{ref}} = 10^4$ (0.738 at 10^3, 0.853 at 10^4, 0.860 at $5 \cdot 10^4$). Configurations where a reference reader leads in Table 1 remain so at every n_{ref}.

With a threshold at the 95th percentile of 1,000 clean ID points and fixed contaminant counts, ID FPR stays at 0.050–0.052 for $\pi \in \{0, 0.005, 0.01, 0.05\}$, and full g_2 exceeds the centered directional readout in AUROC at $\pi = 0.005$ (0.639 vs 0.599), 0.01 (0.677 vs 0.643), and 0.05 (0.774 vs 0.763); with 10–20 OOD points per batch these values are unstable.

D.10 Covariate-shifted ID and mixed OOD

D.10.1 Covariate-shifted ID in the batch

We use ImageNet-1K, both encoders, $n_{\mathrm{ref}} = 5{,}000$, $n = 2{,}000$, $\pi = 0.1$, and five seeds, with the reference and the 1,800 ID points of each batch taken from the Table 1 draws. A fraction ρ of these ID points is replaced by covariate-shifted ID from one of the OpenOOD v1.5 sets ImageNet-V2 (Recht et al., 2019), ImageNet-C (Hendrycks & Dietterich, 2019), and ImageNet-R (Hendrycks et al., 2021), and these points are counted as ID.

ImageNet-V2 has little effect, while with ImageNet-C and ImageNet-R the near-OOD AUROC of g_2 falls below that of Mahalanobis++ by $\rho = 0.5$ and below chance at larger ρ (Figure 6). With

$\rho = 0.5$ and no OOD ($\pi = 0$), a threshold at the 95th percentile of the unshifted ID points flags 65% of ImageNet-C and 84% of ImageNet-R points under g_2, versus about 30% under Mahalanobis++.

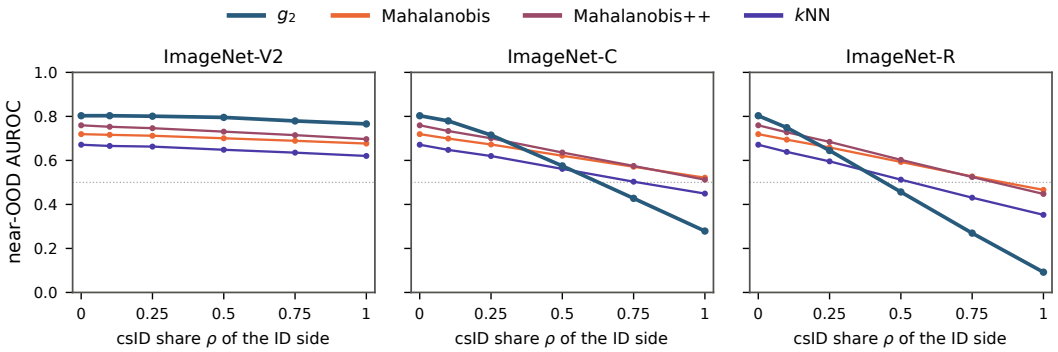


Figure 6: Near-OOD AUROC on ImageNet-1K when a fraction ρ of the ID points in the batch is replaced by covariate-shifted ID (csID) from ImageNet-V2, ImageNet-C, or ImageNet-R, averaged over the two encoders. The dotted line marks chance.

D.10.2 Mixed OOD sources

The setting is that of Appendix D.10.1 without covariate-shifted ID, and the 200 OOD points of each batch are split evenly over near (SSB-hard, NINCO), far (iNaturalist, Textures, OpenImage-O), or all five sources. The ratio readers (g_1, g_2, $\widehat{r}_{\mathrm{kNN}}$) lose 0.012 to 0.096 AUROC relative to the mean of the single-source values in Table 1 (mean 0.055), most on near and all, while the readers of p change by at most 0.014 (mean 0.004). One reading is that several sources spread the batch-reference difference over more directions.